

AHL STUDIES IN THE  
SCIENCE & HISTORY OF LANGUAGE 3

# **Historical Linguistics & Lexicostatistics**

EDITED BY

VITALY SHEVOROSHKIN & PAUL J. SIDWELL

MELBOURNE 1999





**ASSOCIATION FOR THE HISTORY OF LANGUAGE  
MONOGRAPHS AND SERIALS**

MONOGRAPH SERIES 1

AHL STUDIES IN THE  
SCIENCE & HISTORY OF LANGUAGE 3

Editors

Vitaly Shevoroshkin (University of Michigan)  
Paul J. Sidwell (Australian National University)

Volume 3

*Historical Linguistics & Lexicostatistics*



## HISTORICAL LINGUISTICS & LEXICOSTATISTICS

## *Не можем молчать*

Published in Melbourne in 1999  
by the Association for the History of Language

Postal address:  
LPO Box A22  
Australian National University  
ACT, 0200, Australia

Typeset by Fabrice Cavoto (Copenhagen)

Cataloging-in-Publication Data

Shevoroshkin, Vitaly & Paul J. Sidwell (eds.).  
Historical Linguistics & Lexicostatistics

(Association for the History of Language, Monograph Series 1,  
AHL Studies in the Science & History of Language 3)

1. Linguistics--Historical Linguistics--Lexicostatistics--Pre-History. I. Title. II. Series.

ISBN 0 9577251 1 6

© Copyright 1999—Association for the History of Language

\*The views expressed in signed articles are not necessarily the views of the editors or of the  
Association for the History of Language.

# 1st Preface

by V. Shevoroshkin

In the present book several aspects of statistical application for comparative-historical linguistics are investigated; this includes attempts to provide a precise genetic classification of languages. Most papers in the book were written in English for this volume; a few have been translated from Russian.

S. Starostin in his paper on lexicostatistics maintains that it would be incorrect to stop glottochronological studies just because M. Swadesh's methodology shows many inadequacies (one of them being the assumption that a language loses 14 words from the basic 100 over a millennium: the figure seems much too high). One can improve the methodology by both excluding borrowings from the standard diagnostic lists of basic words and amending formulas used in the studies. These formulas should reflect the fact that different basic words of any language have different life expectancies (being, at a certain point in time, replaced by other words with the same meaning).

When properly applied, glottochronology provides, for a given period of time, similar figures for changes in the lexicon both for languages without writing and for languages with an established literary tradition.

Improved glottochronology provides rather precise absolute dating both for separation of any two dialects of a given language, and for separation of any two languages which belong to the given group or family of languages.

Starostin also proposes to use a somewhat different method of glottochronology, namely, root glottochronology (which provides better retention rates when compared with word glottochronology). It is applicable to languages which may be exposed to a reliable etymological analysis in order to determine appropriate roots. Only inherited roots can be chosen for such a study. Among the applications of this method a study of a written text can be used. It is important that the style of the text (colloquial; literary; scholarly) has no bearing on the results.

To illustrate the above method Starostin investigates the 1st page of the original Russian text of his present paper. This text includes 100 roots of Indo-European (IE) origin; they are being compared with appropriate roots in Polish, Lithuanian, German, and French. Using this analysis, as well as comparative data

from similar analyses of other Russian texts, Starostin shows that for every 100 Russian roots one can find in Polish approximately 95 related roots; in Lithuanian—74; in German—54; in French—51. These figures are ultimately determined by the time of the “split” of proto-Slavic into its daughter languages; of proto-Balto-Slavic; of proto-Indo-European.

When taking texts in other languages and comparing them with Russian words which contain appropriate IE roots, Starostin obtains these figures: for 100 German roots—55 Russian; for 100 French roots—50 Russian; for 100 Old Greek roots—52 Russian; for 100 Vedic Sanskrit roots—54 Russian. This means that “the age of the text does not influence the result of the statistical analysis”. Of course, in each case this similarity reflects the distance between proto-Indo-European (of which 100 roots are present in a text written in this or that language: it doesn’t actually matter in what language) and Russian.

For the next experiment Starostin takes the Russian version of the Swadesh 100 word list and compares appropriate IE roots which those retained in Polish (getting 91 Polish cognates), Lithuanian (67—68 matches), German (50—51 matches), and French (50 matches). These results are very similar to those obtained above, when written texts were analysed. Thus, “the distribution of individual frequency characteristics of roots in the Swadesh list coincides with their usual distribution in texts.” Starostin shows here that the dating of language divergence as obtained by word glottochronology practically coincides with the dating obtained when employing root glottochronology. Naturally, this methodology is not applicable if the languages in question are poorly studied.

When using root glottochronology for the study of Nostratic languages, one gets, according to Starostin, 15—20 % rate of root retention for any modern IE, Uralic, Altaic (etc.) language of Nostratic descent: a promising result (the use of a standard 100 word list shows an average 5—9 % word retention in any modern Nostratic languages).

Another paper on lexicostatistics is I. Pejros’ “Family Evolution ...” (Pejros is Starostin’s co-author of the book *Lexicostatistics Revisited*; to appear. Pejros presents parts of the 1st chapter of this book in his paper). Pejros maintains that the Swadesh list is still a useful device of linguistics, though some meanings of the words in this list are not universal: for instance, until recently the meaning ‘horn’ was not known in Australian languages; the meanings ‘bark’ and ‘skin’ are not completely independent; the meaning ‘person’ may be culturally motivated in some languages. Pejros underlines that taboo situations in principle do not impede

glottochronological studies since a speaker is supposed to know both words—the original and the substitute, so the scholar may ignore the latter.

Pejros maintains that neither borrowings nor obsolete words (albeit with good etymologies) may be used as a part of diagnostic word lists. If the meaning of the list is represented by two words (e.g., *palanu* ‘new moon’ and *yilkan* ‘full moon’ in Nyawaygi, an Australian language) both shall be included in the survey. Pejros indicates that on several occasions lexicostatistical analysis was inconclusive because borrowings were included in the diagnostic lists.

Lexicostatistics can be used in language classification: the amount of shared words in any two languages plays here a decisive role. If there are no reconstructions of proto-languages of groups or subgroups, pre-reconstructions can be used,—as they had been indeed used by Pejros for a genetic classification of 66 Australian aboriginal languages (which we had hoped to publish in this volume, but is now slated to appear later in this series).

Pejros’s method of pre-reconstruction in language classification seems to be indirectly confirmed in A. Vovin’s paper on language comparison (published in the present book): Vovin shows that a comparison of basic word lists of two unrelated languages (or, more precisely, languages which belong to different language phyla) does not reveal any meaningful phonetic correspondences, whereas a comparison of two Altaic languages does. Vovin compares 100 Japanese and 100 Turkish words and gets 17 matches; his comparison of Old Japanese and Old Turkish provides 21 matches, and when comparing proto-Turkic and proto-Japanese he obtains 28 reliable matches.—Cf. Starostin’s remark on the degree of relationship of IE languages: “Actually, there is no such IE language which shares less than 20%, and more than 35%, of words with another IE language which doesn’t belong to the same group”.

Two papers in this book—N. Gurevich’s and R. Smiljanić’s—contain critical analysis of the methodology of mass comparison. In the conclusion to his paper Smiljanić writes: “[W]ithout establishing regular sound correspondences, it is hard to rule out similarities due to other reasons than genetic ones and to establish valid groupings which might result in proposing wrong groupings”.—An example which immediately comes to mind is Greenberg’s definition of “Almosan-Keresiouan” family as being part of the Amerind phylum. An analysis which involves traditional methods of comparative linguistics shows that almost all “Almosan-Keresiouan” languages, except Keres (which does show the Amerind pronominal system), are rather closely related to both Na-Dene-Athapaskan

languages of America and Sino-Caucasian languages of Eurasia (cf. my paper on internal and external relationship of Nostratic languages in *Nostratic: Examining a Linguistic Macrofamily*, Cambridge, McDonald Institute for Archaeological Research, 1999, pp. 81–84).

In H. Sverdrup's paper on calculation of language relationship and paths of ancient language/population dispersal in Eurasia, various basic data are used: linguistic, historic, archaeological, biological. As a result, he separates Afro-Asiatic from West Nostratic and East Nostratic, providing a date either around 10,200 BC or 12,000–15,000 BC "for the separation of Afro-Asiatic languages and the split up of Nostratic into different groups" (cf. A. Militarev's and S. Starostin's interpretation of Afro-Asiatic as a sister—not daughter—language of Nostratic). Sverdrup's data indicate an old split between proto-Nostratic and proto-Sino-Caucasian languages (which matches Starostin's comparative data).—He considers Austric as a group of genetically related languages (cf. I. Pejros's historico-linguistic studies) which started to spread around 7,000 BC.

There are many interesting observations also in other papers of this book; most authors agree that improved glottochronology may become a valid tool both for calculating a degree of relationship (genetic distance) between languages and the absolute time of language separation.

*Vitaly Shevoroshkin*  
*Ann Arbor, April 1999*

## 2nd Preface

by P. Sidwell

It is a source of considerable satisfaction that I have been able to help to bring this volume to publication—and even more importantly, to facilitate the continued publication of the occasional series, affectionately known as the ‘red books’, that Vitaly Shevoroshkin inaugurated back in 1987 with *Typology, Relationship & Time*.

‘Red’ no more, but ‘read’ they will continue to be, as the *Association for the History of Language* (AHL—founded in Melbourne in 1992) will henceforth be publishing them as part of its recently founded monograph series *AHL Studies in the Science and History of Language*. Readers who are interested in learning more about AHL will find a potted history of the organisation among the papers in the first AHL monograph *Professing Koernerian Linguistics*, and there is also much of interest available on the AHL homepage at

<http://www.lexicon.net./opoudjis/Work/ahl.html>.

It is truly an important time for the present volume to appear, as lexicostatistics is in grave danger. When the methodology was barely 10 years old, and commensurately still underdeveloped, it was savagely and unfairly criticised by scholars who showed that Swadesh’s estimated rate of vocabulary change gave impossibly recent dates, such as for the separation of Icelandic from Norwegian, and the blow was so great that many were prepared to forget about the matter pretty much from then on. I say ‘unfairly’ for a number of reasons, the most important being that the critics erred seriously in not excluding loan words from their comparisons, failing to consider that the rate of change may be different when loans are so excluded. Instead of drawing insights from their results and suggesting improvements to the pioneering methodology, they were content to throw the baby out with the bath water.

However, while most linguists were, and continue to be, comfortable with the idea that dating languages by lexicostatistical means will always be inaccurate and unreliable, and hesitate to accept the results of such calculations (a view to

which I am also inclined) many continue to use lexicostatistical methods for language classification, and find the results useful and reliable. This is quite understandable, as the main problem with glottochronology, the assumption of constant rate of change, is *utterly irrelevant to genetic classification*.

This point appears to be lost on the most vociferous critics of lexicostatistics, who to this day use every opportunity to rubbish the methodology and belittle its practitioners. For the purposes of genetic classification, the logic is very simple: lexicon changes over time, therefore two related languages will share less as they separate genetically. To put it into a simple slogan, *the greater the distance the greater the difference*. I know of no counter argument to this principle, and I know of no counter examples, i.e. I do not know of anyone claiming that for any real group of languages, those with the highest percentage of commonly inherited vocabulary are the more distantly related, rather than the most closely related to one another.

Remarkably, despite the unassailable logic of this position, there are today senior respected scholars who simply reject the possibility that lexicostatistical methods can offer any useful results for any purpose, and this tendency has the upper hand in the present debate. The loudest of the present crop of naysayers is Professor Lyle Campbell who devotes 10 pages of his otherwise excellent recent textbook *Historical Linguistics: an introduction*, to savaging lexicostatistics. It is regrettable in the extreme that many undergraduate students will find this their first introduction to the concept, and have their views coloured accordingly in their subsequent studies. Campbell misleadingly titles his section *Glottochronology (Lexicostatistics)* and introduces glottochronology as a method of language classification, briefly mentioning that one can distinguish lexicostatistics as a broader term for “the statistical manipulation of lexical material for historical inferences [...] in actual practice, this distinction is almost never made; both names are used interchangeably.” (p.177).

The reality is that scholars who do use lexicostatistical methods are very conscious of the distinction between dating by glottochronology and classifying by lexicostatistics, as these are very different notions. However, having amalgamated both methods into one, Campbell proceeds to take 10 pages to attack them both, listing four ‘basic assumptions’ which have been heavily criticised by scholars since lexicostatistics began.

The first is the *assumption of basic vocabulary*. Campbell states, correctly, that there is no universal, culture-free vocabulary, and that some items on the 100



## PREFACES

word list “seem to be replaced more frequently and more easily than other” (p.183) and claims that these facts are a “problem for the method”. While lexicon may be replaced more or less quickly, and this may be relevant for glottochronological dating, it can only be irrelevant for genetic classification by lexicostatistics. As Starostin (1998) explains:

Within every 100 word list you have a small part, about 20 or 25 words which is really stable and may be unreplaced in 3 or 4 thousand years. Then you have a middle part in which words may be replaced maybe won't, and that is the one which is more significant for classifications like Indo-European or Mon-Khmer etc. And finally there is the least stable part, which accounts for rapid change.

So the more rapidly changing lexicon is important for distinguishing closely related languages and the more stable lexicon is important for higher level groupings—the variable rate of change is actually quite handy and useful. A list of 100 words that don't change, or are exceptionally stable, would actually be less useful!

Campbell's next two assumptions are really one: 2) *constant rate of retention through time*, and 3) *constant rate of loss cross-linguistically*. I agree that these assumptions are proven false, *if borrowings are not excluded from the calculations*. If borrowings are excluded there appears to be an argument that the rate of change may be more stable, and I refer readers to Starostin's paper in this volume. However, as I mentioned above, the rate of change is irrelevant to the degree of genetic distance, so Campbell is misrepresenting the situation to condemn lexicostatistics along with glottochronology on this point. The very logical basis of Campbell's objection to glottochronology is that there can be no fixed relationship between absolute passage of time and relative genetic distance, therefore he must agree that any calculations of relative genetic distance are independent of any rate of change assumptions.

The fourth assumption is *calculation of the date of divergence*, and the above discussion makes this point redundant.

Having attacked glottochronological dating (as it was practiced some decades ago), Campbell then concludes that

For subgrouping, only shared innovations prove reliable, if the cautions about independently occurring changes and possibly inaccurate reconstructions are kept in mind. (p.186)

This advice is all well and good, but the practical reality is that in various language families, such as those of Asia which are morphologically impoverished, it is frequently impossible to identify any subgroupings on the basis of innovations. This is why specialists in these areas continue to use lexicostatistical methods of genetic classification—no other method is available, and there are no convincing arguments as to why such classifications, if they are done with due care to eliminate borrowings, should not be reliable. Until Campbell or another scholar can prove that languages can actually show increased commonly inherited lexicon as they diverge genetically, lexicostatistical methods will continue to enjoy successful, if discrete, application.

The present volume includes substantial contributions from scholars who are at the cutting edge of lexicostatistical practice, and it is hoped that their papers will be read with open minds, and an appreciation for the fact that they reflect developments based upon some decades of experience in working with the methodologies. As scholars we should also be happy to see that we have colleagues who have not been intimidated and silenced by ignorance and intolerance, but carry on in good faith against adversity. While I do not expect readers to be immediately persuaded by all that is presented here, I hope that at least a reasoned debate can now be conducted, and I am honoured to have the opportunity to assist present contributors to present their work in this volume.

Finally I would like to say a special thank you to Fabrice Cavoto and Jim Parkinson for their valuable assistance. Fabrice did a splendid job typesetting most of the papers in this book and liaising with the contributors on many editorial matters (including paving the way for subsequent volumes). Jim prepared the index, so important for making this kind of publication user-friendly. In Melbourne I was assisted by Neile Kirk and Anthony Jukes, both of whom have been consistently helpful. It is only by such valuable volunteer efforts that this book is possible at all, and we should give due regard to those who have given generously of their precious time.

*Paul Sidwell*

*Parkville, October 1999*

## References:

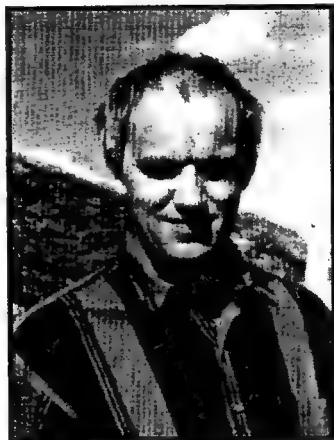
Campbell, Lyle. 1998. *Historical Linguistics: an introduction*. Edinburgh, Edinburgh University Press.

## PREFACES

Starostin, Sergei. 1998. Sergei Starostin on Lexicostatistics & Glottochronology  
*Dhumbadji!: Journal for the History of Language* 4.1:17-21.



Alexis Master-Ramer



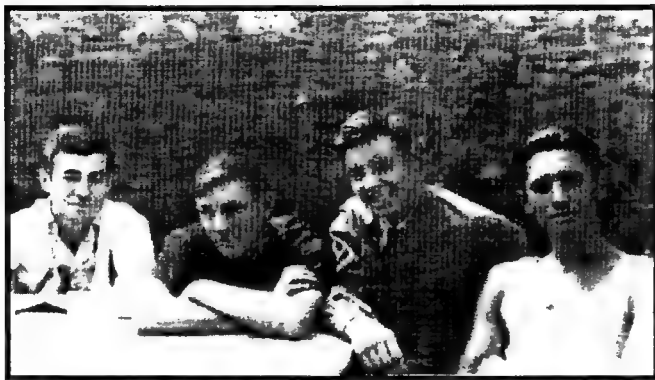
Harald Sverdrup



Ilia Peiros



Sergei Starostin



Vladislav Illich-Svitych (left) with friends, Orenburg, 1951



Vladislav Illich-Svitych (in side-car), 1954



# Contents

Prefaces	v
<b>Part I: Lexicostatistics: Ways of Application</b>	<b>1</b>
Sergei Starostin: HISTORICAL LINGUISTICS AND LEXICOSTATISTICS	3
Sergei Jaxontov: GLOTTOCHRONOLOGY: DIFFICULTIES AND PERSPECTIVES	51
Sergei Starostin: METHODOLOGY OF LONG-RANGE COMPARISON	61
Alexander Vovin: SOME NOTES ON LINGUISTIC COMPARISON	67
Alexis Manaster Ramer: DOLGOPOLSKY'S THEORY OF STABILITY VS. UTO-AZTECAN SECOND PERSON SINGULAR PRONOUNS	95
James Parkinson: THE PRESERVATION OF NOSTRATIC WORD MEANINGS AND SOUNDS, STATISTICAL DATA	105
Henrik Birnbaum: COMMENTS ON SERGEI STAROSTIN'S PAPER ON LINGUISTIC DATING	111
<b>Part II: Genetic Relationship of Languages and "Mass comparison"</b>	<b>117</b>
Naomi Gurevich: PHYLUMPHILE OR PHYLUMFOE? REEXAMINING GREENBERG'S METHOD OF MASS COMPARISON	119
Rajka Smiljanić: WHAT ARE SUFFICIENT CRITERIA FOR ESTABLISHING GENETIC RELATIONSHIPS AMONG LANGUAGES? TESTING 'MASS COMPARISON'	145
<b>Part III: Calculating Language Relationship</b>	<b>167</b>
Harald Sverdrup: CALCULATING LANGUAGE RELATIONSHIPS AND PATHS OF DISPERSAL IN EURASIA DURING THE LAST 100,000 YEARS, USING THE LANGUAGE MODEL	169
Harald Sverdrup and Ramon Guardans: COMPILING WORDS FROM EXTINCT NON-INDOEUROPEAN LANGUAGES IN EUROPE	201
Ilia Pejros FAMILY EVOLUTION, LANGUAGE HISTORY AND GENETIC CLASSIFICATION	259
Indices	307





**Part I:**

**Lexicostatistics: Ways of Application.**



## COMPARATIVE-HISTORICAL LINGUISTICS AND LEXICOSTATISTICS

Sergei Starostin

[This is a translation, done by I. Peiros and N. Evans, of my paper "Sravnitel'no-istoričeskoe jazykoznanie i leksikostatistika", in "Lingvističeskaja rekonstrukcija i drevnejšaja istorija Vostoka", Moscow 1989. I have introduced, however, a number of modifications into the final English text - basically rewritten it again, since the English version needs English examples and etymologies, not Russian ones.]

The last two decades have witnessed a fundamental advance in the techniques of comparative linguistic research. A prolonged period of comparative work with a wide range of language families has laid the foundation for the study of genetic relationships between remotely related languages or language groups. The first step in this direction was taken by V.M. Illič-Svityč in his seminal work 'Towards a comparison of the Nostratic languages' in which, with a combination of rigorous methods and intuitive flare, he begins to demonstrate the relatedness of a number of languages of the Old World.

This new level of comparative studies appears completely legitimate. In fact, if we take the theory of language divergence as axiomatic, we have to concede the fact that from around the sixth millennium B.C. to the first millennium B.C. there was quite a number of different reconstructable proto-languages throughout the world. Once the level of reconstruction of various proto-languages is improved, the question inevitably arises: are any of these proto-languages genetically related and, if so, can we prove this relationship?

To the first part of this question we must now answer in the affirmative. In effect, the absence of genetic relationships between proto-languages (e.g. unrecorded

prehistoric languages) could only be explained by the independent arising of these languages in different parts of the world at approximately the same rather recent period - which would contradict not only simple common sense, but also all our ideas about the history of the human kind and its language. Therefore, at least some existing language families should turn out to be related at an earlier level.

The second part of this question is much more complicated. What is the method of establishing and proving ancient genetic relationships between language families? To this, evidently, there can be only one answer: the classical method of comparative historical linguistics, that is the discovery of a system of regular sound correspondences between proto-languages that is valid for the majority of lexical and morphological items, and the reconstruction of an earlier system based on those correspondences. It is precisely this method that was used by Illič-Svityč in his reconstruction of Common Nostratic.

However, opponents of remote comparison point out that if two languages develop independently for about ten thousand years they inevitably replace practically the entire original lexicon<sup>1</sup>. But ten thousand years or so is exactly the time depth proposed for most macrofamilies. So what can we actually compare in such a case?

Now the situation here is indeed quite complicated, and the probability of establishing the genetic affiliations of language isolates such as Basque, Sumerian or Ainu is extremely small, precisely because of their early separation from any more inclusive language group, and subsequent independent development. But in most cases the situation is different: we have language families and thus the possibility, via the reconstructed proto-languages, of comparing intermediate stages rather than modern

---

<sup>1</sup> In fact this is not entirely the case. The theory of glottochronology maintains that the original 100-word list of the basic lexicon will be entirely replaced in a language over a period of about 15-20 thousand years. If one considers that, in addition to this, any change in the meaning of the word, however small, is regarded by glottochronological theory as the loss of that word, then it becomes clear that the list of roots in the language may be very conservative and that basic roots (perhaps with some semantic modification) can be maintained much longer than words. This motivates my method of root glottochronology, to be discussed below.

languages. Just as we accept that the degree of resemblance between reconstructed proto-Germanic and proto-Slavonic is greater than between modern German and modern Russian, so we must likewise agree that, if proto-Indo-European and proto-Kartvelian are related, then the resemblance between them should be greater than that between modern Russian and Georgian. So the solution, and at the same time the only possible way of dealing with remotely related languages, is in the so-called stepwise reconstruction and comparison of intermediate proto-languages, and the reconstruction from them of an ancestral proto-language.

These are general considerations. In practice, comparativists who study remote relationships face an additional and more serious danger: the possibility of equating genetically unrelated items. These cases can be divided into several categories:

(a) accidental resemblances, such as English *woman* and Old Japanese *womina* 'woman'.

(b) ideophones, such as Russian *kukushka* and English *cuckoo*.

(c) loans, e.g. proto-Japanese *\*kui* and proto-Austronesian *\*kaju* 'tree, wood'.

Not a few works attempting to prove genetic relationships between different language families have foundered just because they failed to take accidental resemblances and old loans into account<sup>2</sup>. The possibility of this kind of mistake is especially great in the case of so-called 'binary' comparisons (which have been particularly popular in American linguistics) - that is, the comparison of two remotely related languages or proto-languages. In fact, if we find in Indo-European a root *\*per-* 'front side' and in Dravidian a root *\*pir-* 'back side', this may well be an accidental resemblance, whose probability could be estimated statistically. But if we add to this pair Altaic *\*p'era* 'lower side', Uralic *\*pera* 'back part' and Kartvelian *\*pir-* 'front side' (see: Illič-Svityč 1971:27) then the probability of coincidental resemblance

---

<sup>2</sup> Critics of remote comparisons sometimes pay too much attention to the second type of case - sound symbolism - as a factor preventing the comparison of remotely related languages. This is the so-called theory of 'elementary relationship'. In my view, however, the role of sound symbolism is exaggerated. It is unclear, for example, why one should think that the deictic stems like *\*mV* 'first person pronoun' or *\*iV* 'demonstrative pronoun' should be regarded as examples of sound symbolism. Therefore I will not discuss this theory in detail.

between all these stems diminishes. If we can find a large number of sets of this kind, demonstrate the regular character of phonological correspondences within them, and discover morphological parallels, then the genetic relationship could be regarded as proven<sup>3</sup>.

Even if we agree that the classical comparative model could be used for the study of remote linguistic relationships, there still remains a host of issues which have not received a proper treatment, but which may be of crucial importance: in particular, the closely interrelated problems of language classification and of dating linguistic divergence.

So - if the Nostratic family exists, then:

(1) What are its branches? For example, are there reasons to divide it into Eastern and Western Nostratic branches? Do Turkic, Mongolian and Tungusic form separate branches of Nostratic, or should they be grouped into a single Altaic branch?

(2) What are its limits? There are some doubts about the correctness of including the Afroasiatic languages in Nostratic, and it is possible that these languages form another macrofamily of approximately the same time depth as Nostratic proper. On the other hand, in some publications on Nostratic there are suggestions that the proto-Nostratic, as reconstructed by Illič-Svityč, could be the proto-language of humankind. How can we prove or reject such statements?

(3) What approximate dates are involved? For example, how could we estimate the time-depth of the different branches of Nostratic? In the literature various estimates for proto-Nostratic range from the ninth to the twelfth millennium B.C or even earlier.

The same questions, of course, could be asked not only about Nostratic but about any other putative macro-family. Moreover, on close examination it turns out that

---

<sup>3</sup> All the above remarks naturally presuppose the classical conception of a 'genetic tree'. If we do not agree with the reality of 'proto-languages', i.e. if we do not think that the reconstructed systems correlate with or are in some sense similar to real languages, then the problem of comparing proto-languages does not arise. Arguments about the problem of remote relationships of languages thus depend on whether one agrees or not with the axioms of historical linguistics, and are thus beyond the scope of this publication.

traditional historical linguistics often gives us contradictory answers even in the case of much younger language families, whose existence is beyond doubt. For example, the limits of the Indo-European family can be regarded as settled, since it is most unlikely that some other known languages could be included in Indo-European, and at the same time it is unlikely that some languages that are regarded as Indo-European do not in fact belong to this family. But the problems of its internal subgrouping and its time depth, as well as its homeland, are still widely debated.

In many cases it is obvious that languages within a family are not equally closely related: one can find many features common to all Slavic languages but not to, for example, Germanic, and these features allow us to assume that Russian is closer to Bulgarian than to Swedish. Evidently if we could measure these distances with the same metric, then we could:

(a) create internal classifications of any language family (or at least evaluate existing classifications) and thus answer questions about internal subgrouping.

(b) locate a language or language family within higher-level units such as families or macro-families and thus answer questions about the limits of families.

(c) combine this metric with a time-scale to estimate the probable time-depth of linguistic divergence, thus answering questions about chronology.

It is thus of central importance to find quantifiable characteristics of languages that can be used as objective criteria in establishing distances between languages<sup>4</sup>. Measurements based on phonological, morphological or syntactic similarities obviously can not be used as such criteria, since such systems are known to change rapidly and radically in some cases, and to be extremely conservative in others. The rate of change in these areas is thus uneven and in any case it is not obvious that it could be measured: a statistical measurement of their rate of change is hindered by the small number of elements (phonemes, grammatical morphemes), which prevents us from obtaining sets that are sufficiently populous for statistical tests<sup>5</sup>.

---

<sup>4</sup> It is obvious that such criteria could also be used to verify the relationship between languages, and thus the distance between two unrelated languages should be estimated as infinite.

<sup>5</sup> Of course it does not follow from this that phonological and morphological data cannot be used in

The one domain of language in which the rate of change appears to be nearly regular (see below) and which is at the same time an appropriate object for quantitative statistical measures is the lexicon. Therefore I would like to discuss, in the remainder of this paper, various aspects of lexicostatistics and glottochronology.

Despite a number of critical works showing many inadequacies of the glottochronological method invented by Swadesh in the 1950s (to the point where it seemed at one time that glottochronology had been fully discredited), I think it would be not only premature to abandon it, but also incorrect. Every comparativist who has worked with glottochronology knows that closely related dialects usually have a cognacy rate of 90% or more on Swadesh's 100-word list: closely related languages (such as those within the Slavic, Romance, Germanic or Turkic groups - that is those which diverged around one and a half to two thousand years ago) share from 70 to 80% of items on this list, and language families such as Indo-European which split up five or six thousand years ago have a rate of 25 to 30%. Once we start to talk about more ancient families such as Uralic or Altaic we find a rate of at most 10 or 20 percent. Finally, the cognacy rate for modern languages belonging to different branches of such a macro-family as Nostratic is even less - around 5 - 9 %.

These figures are, to be sure, approximate, and more precise data will be given below. But one should note that at each level the cognacy rate is systematically replicated: no Slavic language has a cognate rate with any other Slavonic language of less than 75%; no Indo-European language has a rate of more than 35% or less than 20% with any other Indo-European language of a different branch. It would be quite absurd to obtain a figure, say, of 50% between Russian and Polish, or Russian and German. It is well-known that similar rates of cognacy are found between languages related at an equivalent level in other language families, such as Austronesian, Uralic, Sino-Tibetan etc. All these considerations give us some indication that the rate of

---

the genetic classification of languages, since it is well-known that most existing classifications are based on such data. I am only stating that such data are insufficient for quantitative evaluations of the distances between languages. Thus we can separate the Min dialects from all other Chinese dialects by the fact that they maintain the oppositions *\*b* - *\*bh*, *\*d* - *\*dh* etc., which are lost in all other dialects. But this feature does not help us estimate the depth of their separation.



change in the lexicon (in some of its domains at least) really might be steady and universal. There is thus reason to discuss once again the fundamental postulates and methodologies of glottochronology.

The fundamental principles of Swadesh's glottochronology could be represented by five postulates, well-formulated in Arapov and Herz (1974):

" 1. In the lexicon of any language one can distinguish a particular portion which we will call basic or stable.

2. One can provide a list of meanings which in any language of the world will be represented by words from its basic vocabulary...We shall say that these words form the basic list, BL. Let us represent the number of words in BL by  $N_0$ .

3. The proportion  $P$  of words from BL which remain (i.e. are not replaced by other words) over a time interval  $t$  ... is constant. That is, it depends only on the amount of time elapsed, and not on how that interval was chosen, or on which words from what language are considered.

4. All words from BL are equally likely to be retained or not to be retained during a particular period of time.

5. The probability of a word from a proto-language's BL being retained in the BL of one of its daughter languages is independent of its probability of being maintained in the comparable list of any other daughter language." (ibid: 21-5)

All the above postulates are used to obtain the basic mathematical equation of glottochronology:

$$(1) N(t) = N_0 e^{-\lambda t}$$

where the time elapsed between two points of development is denoted by  $t$  and is measured in millennia;  $N_0$  is the initial BL,  $\lambda$  is the 'rate of loss' of words from  $N_0$ ; and  $N(t)$  is the proportion of words from the initial list retained at time  $t$ . Thus, if the original list includes one hundred words and is regarded as 1, and if  $\lambda = 0.14$  (i.e. during one millennium fourteen out of one hundred words will be lost), and time elapsed from the beginning of dispersion is, for example, two thousand years, then

$$N(2) = e^{-0.14 \times 2} = 0.76 \text{ (i.e. seventy-six words).}$$

Knowing the proportion of words from BL retained in a given language we can calculate, with the help of logarithms the period of elapsed time:

$$(2) t = \frac{\ln N(t)}{-\lambda * N}$$

According to the fifth postulate we assume that the development of two languages from one proto-language was independent. Thus, knowing the cognacy rate in BL lists in two or more related languages, we can work out the time since their separation. The retention rate between  $n$  languages, which have a common ancestor with a single BL, is:

$$(3) N_n(t) = N_0 e^{-n\lambda t}$$

and the time of their separation can be obtained by the formula

$$(4) t = \frac{\ln N_0(t)}{-n\lambda N_0}$$

Thus, if  $\lambda = 0.14$  as above, and the retention rate between two languages is 0.8 (i.e. 80%), then the time since separation is

$$t = \frac{\ln 0.8}{-2 * 0.14} = 0.8 \text{ (i.e. about 800 years)}^6$$

One should note in particular that all datings so obtained are probabilistic rather than absolute in character, and allow the possibility of errors of various magnitudes, whose confidence interval may theoretically be calculated, but a discussion of this is

---

<sup>6</sup> In publications on glottochronology one often meets a somewhat simplified notation of the mathematical correlations discussed here, where instead of 'rate' one uses the coefficient of lexical retention  $r = 1 - \lambda$ . The following notational variants of the formula are frequently encountered:

(1) as (1')  $c = r^t$  where  $c$  corresponds to  $N(t)$ , and  $N_0$  is assumed to be 1.

(2) as (2')  $t = \frac{\log c}{\log r}$

(3) as (3')  $c = r^{nt}$ , or more often as  $c = r^{2t}$  where  $n = 2$ .

(4) as (4')  $t = \frac{\log c}{n \log r}$ , again more often encountered

as  $t = \frac{\log c}{2 \log r}$ , for  $n = 2$ .

beyond the scope of this article.

The 'rate'  $\lambda$  is a constant, which has been established empirically on the basis of samples from languages with a long history recorded over the last millennium or more. The value of 0.14 for  $\lambda$ , which was used above, is not arbitrary: precisely this constant was postulated by the founding figure of glottochronology, Morris Swadesh, for his one hundred word BL (see, e.g., Swadesh 1960).

The above theory of glottochronology (whose mathematical apparatus is practically borrowed from the physical theory of radioactive split) is rather elegant and simple. Unfortunately, however, we must hasten to point out that for linguistic datings, Swadesh's version of glottochronology is inappropriate<sup>7</sup>. In practice, in all cases of historically recorded linguistic events we find one and the same outcome: all dates given by standard glottochronology are much younger than the historical records suggest. Let us try to discuss the reasons for this phenomenon.

First of all we should discuss the question of whether the retention rate of BL is in fact constant (recall that this is the third postulate of glottochronology). This question has been the subject of wide debate, and it is obvious that in the absence of a constant rate the whole procedure of glottochronology becomes senseless.

In the literature on glottochronology one can find the claim that 'the assumption about changes in the lexicon does not apply to languages with an established literary tradition' (see Jaxontov 1984:45). This position held by supporters of glottochronology is a reaction to criticisms put forward in the classic article by Bergsland & Vogt (1962). These scholars, in discussing Scandinavian material, have shown that the rate of change in the BL for Icelandic over the last millennium was only about 0.04 (retention rate  $r = 0.96$ ) while in literary Norwegian (Riksmål) it was about 0.2 ( $r = 0.8$ ). Accordingly we obtain, using Swadesh's value of 0.14 for the constant, improbable results: from 100 to 150 years for Icelandic and 1400 years for Riksmål, despite the fact that both languages developed from the same source,

---

<sup>7</sup> Perhaps this is the reason why many scholars, using the glottochronological method no longer try to use it for absolute datings, but only for relative datings (i.e. for creating genetic trees).

independently, over about 1000 years. Comparable results were obtained by O'Neil (1964) from a comparison of Icelandic and Faroese BLs: the dating of their divergence is known historically (C10 A.D.), but the 94% of shared lexicon between these languages suggests, according to Swadesh's formula, that the divergence started 200 years ago.

These are, it would seem, obvious facts, and they forced scholars to make allowances for the effects of a literary norm which hinders the development of the lexicon - though in the case of Riksmål, as we have seen, we have on the contrary a more rapid development. Let us try, however, to discuss the Scandinavian case.

It is easier to explain the accelerated development of Riksmål. This language is actually a hybrid of Norwegian and Danish, and its one hundred word list includes eleven Danish, three Swedish and two German loans. If we treat all these loans as replacements in our comparison with Old Icelandic then we will get a rapid rate of development for Norwegian. On the other hand, if we exclude these borrowings from consideration we get a rate for Icelandic equal to 0.06, with six words having changed: *eta* 'eat' > *borða*, *lauf* 'leaf' > *blað*, *verr* 'man' > *maður*, *swima* 'swim' > *synda*, *varmr* 'warm' > *hlyr*, *mani* 'moon' > *tungl*. For Riksmål the rate is 0.05, with four words replaced - *eldr* 'fire' > *varme*, *lauf* 'leaf' > *blad*, *hold* 'meat' > *kjött* and *sa* 'that' > *den* - from the list of 84 original words. A very similar rate will be obtained for other Scandinavian languages: 0.05 for Faroese, and 0.04 for Swedish, Danish and the Norwegian dialects Gwestal and Sandnes<sup>8</sup>.

Thus we can see that if we regard borrowings as lexical replacement we get serious errors. This leads to an important conclusion: before doing glottochronological calculations one should eliminate all borrowings from the BL list (at least where there have been intensive borrowings between two neighbouring languages), paying attention only to the rate of change within non-borrowed lexicon<sup>9</sup>.

<sup>8</sup> These values can vary somewhat (from 0.04 to 0.06) depending on which words for 'earth' (*mold* or *jord*) and 'meat' (*hold* or *kjöt*) we choose as the main words in Old Icelandic. This has little effect on the overall results.

<sup>9</sup> We thus regard changes in the original lexicon as a regular process, but the substitution of original

If we do so, the rate of development of the BL lexicon in Scandinavian languages ends up being stable and not significantly different from 0.05. It is clear, however, that this speed is much slower than the constant of 0.14 postulated by Swadesh. So it looks as if we should follow Bergsland and Vogt in assuming a slower rate of replacement in the Scandinavian languages compared to others. Now let us turn to other languages. We have at our disposal data on the development of the BL in many well-documented languages of the world, which give the following picture:

Language	t	$\lambda_1$	$\lambda_2$
Japanese	1.2	0.11	0.06 <sup>10</sup>
Chinese	2.6	0.1	0.1 <sup>11</sup>

words by loans as chance mutations that require additional correction. From this it immediately follows that glottochronological calculations become worthless in conditions where the historical and etymological situation is not sufficiently well understood for us to separate original words from loans. I would like to emphasize that no lexicostatistical survey is possible until thorough comparative work has been done. Thus, for example, the lexicostatistical calculations by Shiro Hattori, who compared the Japanese 100-word list with lists of many different languages of the world (Hattori 1959), or the calculations of Swadesh himself, who has measured the level of lexicostatistical similarity between different languages of Eurasia and America (the "Dene-Finnish" theory of Swadesh 1965), are obviously senseless.

The importance of dealing with loanwords in lexicostatistics was stressed in a recent study by Sheila Embleton (Embleton 1986), who also attempts to introduce some corrective coefficients in cases of language contacts. This is an excellent book, proving that lexicostatistics is still very much alive.

<sup>10</sup> In modern Japanese, in comparison to old Japanese (eighth century A.D.) the following words have been changed: *kap-u* 'eat' > *taberu*, *kapigwo* 'egg' > *tamago*, *kasira* 'head' > *atama*, *isagwo* (managwo) 'sand' > *suna*, *wi-ru* 'sit' > *suwaru*, *ka-no* 'that' > *ano*, *nare* 'you' > *anata* (*kimi*); and the following words have been borrowed: *kokoro* 'heart' > *shinzō*, *kimo* 'liver' > *kanzō*, *sisi* 'meat' > *niku*, and *kapa* 'skin' > *hifu*, *pi* 'sun' > *taiyō*.

<sup>11</sup> In modern Chinese, in comparison with Archaic Chinese (7th century B.C.), with the total absence of loans, the following words have been replaced: *krej* 'all' > *yiqie*, *duo*; *puk* 'belly' > *tuzi*; *crū?* 'fingernail' > *jia*; *khwīn* 'dog' > *gou*, *ʔem* 'drink' > *he*; *lek* 'eat' > *chi*; *rōn* 'egg' > *dan*; *cok* 'foot' > *jiao*; *pic* 'give' > *gei*; *raŋ* 'good' > *hao*; *keŋ* 'neck' > *bozi*; *\*khek* 'red' > *hong*; *wat* 'say' > *shuo*; *kēnh* 'see' > *kan*; *rəp* 'stand' > *than*; *\*nī* 'sun' > *taiyang*; *cə, də?* 'this' > *zhe*; *slhaj, paj?* 'that' > *na*; *thə?* 'tooth' > *ya*; *mōk* 'tree' > *shu*; *níc* 'two' > *liang*; *grāj* 'go' > *zou*; *ghāj* 'what' > *shemna*. Unfortunately it is rather difficult to compile lists for intermediate stages of Chinese, since we have an accurate representation of the spoken language only in ancient texts on the one hand, and in the modern language on the other.

English	1.3	0.14	0.1 <sup>12</sup>
German	1.2	0.08	0.05 <sup>13</sup>
French	1.5	0.09	0.07 <sup>14</sup>
Spanish	1.5	0.07	0.06 <sup>15</sup>
Rumanian	1.5	0.09	0.06 <sup>16</sup>

This list could be enlarged: in particular it has been shown by Fodor (1961) that there has been the same low rate of development in the lexicon of the Slavonic languages. It is clear, however, that the rate of development of BL lists over the last one to one and a half millennia is much less than the Swadesh constant of 0.14, which

<sup>12</sup> In modern English the following words have been changed compared to Old English of the ninth century A.D.: *wamb* "belly", *micel* "big", *fuyol* "bird", *wolcen* "cloud", *hund* "dog" (perhaps a Low German loan), *yuma* (*wer*) "man", *flæsc* "meat", *heals* "neck", *wey* "road", *rec* "smoke", *se* "that", *þu* "you". The loans are *rinde* > bark, *steorfan* > die, *beory* > mountain, *sinwealt* > round, *hyd* > skin.

<sup>13</sup> In modern German, compared to Old High German of the 8<sup>th</sup> century A.D., the following words have been changed: *bein* "bone" > *Knochen*, *wamba* "belly" > *Bauch*, *mihhil* "big" > *groß*, *luzzil* "small" > *klein*, *zagal* "tail" > *Schwanz*, *wib* "woman" > *Frau*. The loans are *feizzit* "fat" > *Fett*, *houbit* "head" > *Kopf* and *sinwel* "round" > *rund*.

<sup>14</sup> In Modern French, in comparison to Vulgar Latin of the 4th or 5th Century A.D., the following words have been changed: *penna* "feather" > *plume*, *caput* "head" > *tête*, *audire* "hear" > *entendre*, *occidere* "kill" > *tuer*, *multum* "many" > *beaucoup*, *carnem* "meat" > *viande*, *arenam* "sand" > *sable*, *sementem* "seed" > *graine*, *ambulare* "go" > *aller* (possibly a loan), *natare* "swim" > *nager*, *parvus* "small" > *petit*; and *albus* "white" > *blanc* is a loan.

<sup>15</sup> In Spanish, compared to Vulgar Latin of the 4th to 5th century A.D., the following words have been replaced: *canem* "dog" > *perro* (possibly a loan), *occidere* "kill" > *matar*, *genuculum* "knee" > *rodilla*, *stare collocatum* "lie" > *estar acostado*, *longus* "long" > *largo*, *parvus* "small" > *pequeño*, *ambulare* "go" > *andar*, *galbinus* "yellow" > *amarillo*. The loans are *pennam* "feather" > *pluma* and *albus* "white" > *blanco*.

<sup>16</sup> In Modern Rumanian, in comparison to Vulgar Latin of the 4th or 5th Century A.D., the following words have been changed: *ventrem* "belly" > *pontece*, *grandem* "big" > *mare*, *avem* "bird" > *pasăre*, *ustulare* (cremare) "burn" > *arde*, *frigidus* "cold" > *rece*, *siccus* "dry" > *uscat*, *terram* "earth" > *pământ*, *cordem* "heart" > *inima*, *buccam* "mouth" > *gura*. Loans are *collum* "neck" > *gât*, *camminum* "road" > *drum*, *arenam* "sand" > *nisip* and *parvus* "small" > *mic*.

can be obtained only for English and only if one treats loans as replacements<sup>17</sup>. We will defer for the moment our explanation of the comparatively high rate of development (0.1) of the Chinese lexicon.

Looking at the data above, it would seem easy to change Swadesh's value of 0.14 for  $\lambda$  to 0.06, which is the average of all  $\lambda_2$  in the above cases. Taking this value of 0.06 will give dates for the Scandinavian languages which are slightly too young, and more or less exact datings for all the Romance languages, and for the remaining Germanic languages other than English. But we will get an absolutely unrealistic dating for Chinese (i.e. middle of third millennium B.C.), and the dating for Old English will be one thousand years too early. Moreover, if we use such a value of  $\lambda$  in formula 4 constant to date the divergences of various languages, the result is a fiasco. Thus, for the disintegration of Belorussian and Ukrainian (which share 97% on their 100 list) we obtain:

$$t = \frac{\ln 0.97}{-2 * 0.06} = 0.25$$

That is, a separation 250 years ago, a date which obviously corresponds to nothing<sup>18</sup>. The same period of independent development would separate, for example, modern Persian and Tadjik, Yiddish and German, and so on. The datings here turn out to be too young. On the other hand, if we go back to the Indo-European level and try to give dates, for example, for the period of separate development of Russian and Persian (which share 28% on a one-hundred word list), we obtain:

---

<sup>17</sup> In the remaining languages, even if we count the loaned items, the rate is substantially less than 0.14. The somewhat 'accelerated' development of English should be probably explained by active language interference during the formation of Modern English, with Scandinavian and Romance components playing a part.

<sup>18</sup> For the separation of Belorussian and Ukrainian we obtain a date of the mid nineteenth century using the Swadesh formula; for the separation of Icelandic and Faroese, as mentioned above, the eighteenth century. For the disintegration of proto-Slavic, of proto-Germanic, and proto Romance (i.e. Vulgar Latin) we obtain datings around the 11<sup>th</sup> or 12<sup>th</sup> centuries A.D., and for the disintegration of Common Indo-European, the middle or end of the third millennium B.C.

$$t - \frac{\ln 0.28}{-2 \cdot 0.06} = 10.6$$

- that is, the beginning of the ninth millennium B.C. But the more or less established view is that the disintegration of common Indo-European took place in the fourth millennium B.C. Other pairs of Indo-European languages will also yield datings that are far too early.

Now we can understand why Swadesh chose a value of 0.14 for  $\lambda$ . With a normal exponential correlation between time and percentage retention this value avoids giving us unreasonably old dates at great time depths, while at the same time giving slightly too recent dates for the whole period with externally verifiable datings. We have already seen that the value of 0.14 is in reality found only in English, and even there some caveats regarding borrowings are needed. It follows that it cannot have the status of an empirically established value. On the other hand, the empirically observed value of  $\lambda$  allows us to give reasonable datings for events, which happened one millennium or one and a half millennia ago, but is absolutely unsuitable for dating earlier or later splits. We need to understand the nature of this contradiction and to find some kind of solution.

Let us once more turn to the case of the Romance languages. We estimated above the rate of replacement on the basis of a comparison of the BL lists of the Romance languages with the BL of Vulgar Latin, the proto-language of all modern Romance languages which disintegrated around the fourth to sixth centuries A.D. Let us now compare the BL of classical Latin, which dates from the fourth to second centuries B.C. and has a well-known BL<sup>19</sup>, with a view to what changes took place in the French, Spanish and Rumanian BLs in comparison with classical Latin.

The following words have been changed in French: *omnis* > *tout* 'all', *magnus*

---

<sup>19</sup> Vulgar Latin is dated differently by different scholars (cf. Müller 1929, Guryčeva 1959). But it is obvious that it had a unitary nature until the fifth century A.D. despite the presence of some dialectal differences. It began to separate into dialects between the fifth and eighth centuries A.D., and the period of Romance languages dates back to the eighth century.



> *grand* 'big', *edere* > *manger* 'eat', *pinguedo (adepts)* > *graisse* 'fat', *penna* > *plume* 'feather', *ignis* > *feu* 'fire', *caput* > *tête* 'head', *audire* > *entendre* 'hear', *occidere* > *tuer* 'kill', *cubare* > *être couché* 'lie', *jecur* > *foie* 'liver', *vir* > *homme* 'man', *multus* > *beaucoup* 'many', *caro* > *viande* 'meat', *os* > *bouche* 'rot', *arena* > *sable* 'sand', *semen* > *graine* 'seed', *cutis* > *peau* 'skin', *parvus* > *petit* 'small', *nare (natate)* > *nager* 'swim', *ire* > *aller* 'go', *flavus* > *jaune* 'yellow'. The following are loans: *via* > *chemin* 'road', *lapis* > *Pierre* 'stone' and *albus* > *blanc* 'white'. We thus have a 77% retention of lexicon; letting  $t \approx 2.3$  we get a value of  $\lambda t = 0.11$ , whereas the rate between Vulgar Latin and Modern French is  $\lambda = 0.07$  - see above.

In Spanish the following words have changed: *omnis* > *todo* 'all', *magnus* > *grande* 'big', *urere* > *quemar* 'burn', *canis* > *perro* 'dog', *pinguedo (adepts)* > *grasa* 'fat', *ignis* > *fuego* 'fire', *occidere* > *matar* 'kill', *genu* > *rodilla* 'knee', *cubare* > *ester acostado* 'lie', *jecur* > *higado* 'liver', *longus* > *largo* 'long', *vir* > *hombre* 'man', *os* > *boca* 'mouth', *via* > *camino* 'road', *cutis* > *piel* 'skin', *parvus* > *pequeño* 'small', *ire* > *andar* 'go', *flavus (fulvus)* > *amarillo* 'yellow'. The following are loans: *penna* > *pluma* 'feather', *lapis* > *piedra* 'stone', and *albus* > *blanco* 'white'. We thus have a retention rate of 80%; letting  $t = 2.3$  we get a value of  $\lambda t = 0.1$ , whereas the rate between Vulgar Latin and Modern Spanish is  $\lambda = 0.06$  - see above.

In Rumanian the following words have changed: *omnis* > *tout* 'all', *venter* > *potece* 'belly', *magnus* > *mare* 'big', *avis* > *pasăre* 'bird', *urere* > *arde* 'burn', *frigidus* > *rece* 'cold', *siccus* > *uscat* 'dry', *terra* > *pămînt* 'earth', *edere* > *monca* 'eat', *pinguedo (adepts)* > *grasime* 'fat', *ignis* > *foc* 'fire', *cor* > *inimă* 'heart', *cubare* > *sta culcat* 'lie', *jecur* > *ficat* 'liver', *os* > *gură* 'mouth', *vir* > *om (barbat)* 'man', *cutis* > *piele* 'skin', *ire* > *umbla* 'go', *flavus (fulvus)* > *galben* 'yellow'. The loans are: *collum* > *got* 'neck', *via* > *drum* 'road', *arena* > *nisip* 'sand', *parvus* > *mic* 'small', and *lapis* > *piatră* 'stone'. This is the same retention rate of 80%, as Spanish, which gives us the same rate over 2.3 thousand years of  $\lambda t = 0.1$ , again compared to

the rate of  $\lambda = 0.06$  between Vulgar Latin and Rumanian as shown above.

We see, then, a clear case of the rate of change increasing (i.e. an acceleration) with an increased rate of lexical loss in BL as the separation time  $\lambda t$  increases. This fact explains the extremely low rate of lexical loss since the separation of Belorussian and Ukrainian, and also between Persian and Tadzhik: we get 97% of correspondences between each of these pairs, both of which separated about 600 years ago, and thus  $\lambda$  should be set at about 0.03 for these cases<sup>20</sup>. At the same time, this is the explanation of the comparatively high figure of 0.1 for the rate of loss within the Chinese lexicon: the time elapsed between Archaic and Modern Chinese is about the same as that between Classical Latin and the modern Romance languages.

As mentioned above, the mathematical apparatus of glottochronology was borrowed from the theory of radioactive decay. But the most important difference between words and neutrons is perhaps the fact that the former, by contrast with the latter, can become 'older'. In fact, the probability of a neutron remaining intact at a given time is always  $e^{-\lambda t}$ , regardless of its history. On the other hand, the probability of a word (including those in BL) remaining correlates with how long this word has already 'lived'.

We can thus explain, for example, the replacement in all modern Indo-European languages of the proto Indo-European word *\*g<sup>w</sup>(e)ru-* 'heavy': this word is represented in Vedic *guru-*, Ancient Greek *βαρύς*, Lat. *gravis*, Gothic *kaurus*, but is lost in the majority of modern languages: with the meaning 'heavy' this stem is maintained only in Modern Greek and in some modern Indian languages, while in most Indian, Germanic, Romance, Slavonic and other Indo-European languages it has either changed its meaning or disappeared. Comparativists are familiar with many examples of this kind -namely the wide distribution of some words or roots in ancient languages

---

<sup>20</sup> The separation of Belorussian and Ukrainian can be dated by historical records to the thirteenth or fourteenth centuries, if one correlates this with the separation of Belorussia after its conquest by Lithuania. For the separation of Persian and Tadzhik the probable dating is the end of the fourteenth or the beginning of the fifteenth century, when Tadzhikistan and the eastern part of Iran separated as a result of the Mongolian invasion.

and their almost total absence from the modern languages of the same family - which seem to be connected to the 'lifetime' of given words.

The suggestion that words can 'age' automatically leads us to reject the third postulate of glottochronology - that is, of the idea that the rate of retention in BL is constant - and to adopt the hypothesis that there is a correlation between  $\lambda$  and time  $t$ . And in fact if we assume that the greater the value of time  $t$  then the higher the probability that an original word from BL will disappear, then instead of (1) we adopt as our main formula in glottochronology

$$(5) N(t) = N_0 e^{-\lambda t^2}$$

This is the formula for regular acceleration of speed<sup>21</sup>.

Now, the time  $t$  can be established by the following formula:

$$(6) t = \sqrt{\frac{\ln N(t)}{-\lambda * N_0}}$$

and the time of separation between  $n$  languages by the formula

$$(7) t = \sqrt{\frac{\ln Nn(t)}{-n\lambda N_0}}$$

If we insert the data given above for the various languages in this formula we obtain the coefficient of acceleration  $\lambda$ :

Language	T	$\lambda$
Japanese	1.2	0.05
English	1.3	0.08
Chinese	2.6	0.04
German	1.2	0.04

<sup>21</sup> The possibility of such a correlation was already envisaged, though only theoretically, in Arapov & Herz 1974.

French	1.5	0.05
Spanish	1	0.06
Riksmal	1	0.05

With the exception of English, which reveals a somewhat higher value, the value for  $\lambda$  is stable and varies only slightly between 0.04 and 0.06. In addition, when we date language separation over the last thousand to one and a half thousand years the above formulae on the whole give good results. Thus for Belorussian and Ukrainian as well as for Persian and Tadzhik we have:

$$t = \sqrt{\frac{\ln 0.97}{-2 \cdot 0.05}} = 0.55$$

(i.e. the end of the fourteenth century A.D.)

For the date of separation of Icelandic and Faroese (94% correspondence rate) we get about the twelfth century instead of the eighteenth century according to the Swadesh formula, and so on.

However, one can show that the use of formula (5), when applied to separations of greater time depth, begins to yield dates that are too recent. Thus for Russian and Persian the application of (7) with a value of 0.05 for  $\lambda$ , the time of separation would be

$$t = \sqrt{\frac{\ln 0.28}{-2 \cdot 0.05}} = 3.6$$

i.e. around the middle of the second millennium B.C. Even with the help of the Swadesh formula, with 28% shared lexicon we will obtain the end of the third millennium B.C. whereas the real dating should rather be the fourth millennium B.C. (see above).

Moreover, one can show that if we adopt formula (5) the period for the complete replacement of the BL should be less than 10,000 years, and for any two separated languages all correspondences on the one hundred word list should have disappeared

after seven millennia. There is no doubt that, if we adopt the formula

$$(5) \quad N(t) = N_0 e^{-\lambda t^2}$$

the dating will be somewhat improved for recent periods, and will agree better with real instances of lexical development. But for more ancient periods it will give results that are worse than those of Swadesh. How do we resolve this contradiction?

In order to advance further, we need to discuss the fourth postulate of glottochronology: 'all words of BL have the same chances of being retained over a given  $t$ ' (above). It was this statement that aroused the strongest objections both among supporters of glottochronology and its detractors. In fact, any linguist who works with glottochronology knows that not all words in BL are of equal stability. The words 'small' or 'skin' have on the whole a better chance of disappearing from the list than such words as 'T', 'you' or 'ear'.

Therefore it has often been suggested that the coefficient of the retention rate for the whole list should in general be derived from the individual coefficients of retention for each word, i.e. the probability of it being retained over a given period of time  $t$  (see, e.g. van der Merwe 1966). An attempt to determine such individual coefficients empirically (for Austronesian and Indo-European languages) has been made by Dyen and James (1967), but their mathematical apparatus was too baroque and unproductive. Moreover, if the suggestion that different words have different retention rates is true, it is also quite probable that these individual coefficients of retention could vary according to the cultural and linguistic environment. For example, such words as 'cloud' and 'tail' are very stable in the Turkic languages but unstable in Germanic; the word 'belly' is very stable in Romance but unstable in Slavic and so forth. Although some words from the BL do indeed reveal a high level of stability ('T', 'you', 'sun', 'eye', 'ear' etc.), it is clear that a formula based on individual coefficients cannot solve this problem, because it will not be universally applicable.

Let us imagine now an ideal BL list with some average rate of divergence  $\lambda$ , where all words are ranked according to the probability of their disappearance over a given period of time  $\Delta t$ , with the first word in the list having a probability of being

retained close or equal to zero, while the last word has a probability approaching 1. In this case, words should disappear in turn, beginning with the least stable and going on to the more stable. Accordingly, at time  $t_n$ , the rate of loss for those items remaining on the list should become slower than at time  $t_{n-1}$ . The rate of loss of this list will, it follows, be variable, depending at any moment on the proportion of retained (and therefore of lost) words:

$$(8) \lambda_{t_n} = \lambda_0 \cdot N(t_n)$$

If we put (8) in formula (5) we obtain the formula:

$$(9) N(t) = N_0 e^{-\lambda N(t)t}$$

Time (t) can be calculated according to the formula:

$$(10) t = \sqrt{\frac{\ln N(t)}{-\lambda \cdot N(t) \cdot N_0}}$$

Note that in the case where we have n languages, the formula giving the date of their disintegration can be recast as:

$$(11) N_n(t) = N_0 e^{-n\lambda N(t)t}$$

where  $N(t)$  is the value, in reality not attested, but obtainable more or less precisely by the formula:

$$(12) N(t) = \sqrt[n]{N_n(t)}$$

From (11) and (12) we get a means of calculating the time of divergence for a proportion of corresponding words in the BL of n languages:

$$(13)^{22} \quad t = \sqrt{\frac{\ln \left( \frac{Nn(t)}{N_0} \right)}{-n\lambda^n \sqrt{Nn(t)}}$$

Formula (9) represents the 'contradictory' character of the process of lexical loss in BL: the square of  $t$  reflects the acceleration of loss caused by the 'aging' of words, while the coefficient  $N(t)$  in the exponent reflects the opposite to this - the deceleration of the rate of loss when less stable words disappear from the remaining BL list and the more stable items are retained<sup>23</sup>. We may note that for small values of  $t$  (and accordingly for large  $N(t)$ ), the datings according to formulae (9) and (5) will be similar. In contrast, as  $t$  grows (and  $N(t)$  diminishes) the datings become significantly earlier. If we assume that  $\lambda$  is 0.05 on average (see above), we get the following table of selected datings, where  $N(t)$  is the proportion of words from BL remaining in one language,  $N_2(t)$  is the proportion of words corresponding between two languages,  $t$  is the time of development and therefore of divergence in millennia:

$N(t)$	$N_2(t)$	$t$	$t$ (According to M. Swadesh)
0.99	0.99	0.3	0.03
0.97	0.94	0.8	0.2
0.95	0.9	1	0.35
0.9	0.81	1.5	0.7
0.85	0.72	2	1.1
0.8	0.64	2.4	1.5
0.75	0.56	2.8	1.9

<sup>22</sup> For practical purposes (in the most frequent case, when  $n=2$ ,  $N_0=1$  (100 words) with  $\lambda = 0.05$  and with a more traditional rendering of  $N(t)$  as  $c$ ) the formula may be simplified as:

$$t = \sqrt{\frac{\ln c}{-0.1 \sqrt{c}}}$$

<sup>23</sup> A thorough criticism of the mathematical apparatus of glottochronology was undertaken by Chréien (1952). However all his objections were effectively countered by Dyen. We will therefore not discuss Chréien's objections here, but simply note that they have no connection with the problem of the variability of retention rates.

0.7	0.49	3.2	2.4
0.65	0.42	3.7	2.9
0.6	0.36	4.1	3.4
0.55	0.3	4.7	4
0.5	0.25	5.3	4.6
0.45	0.2	6	5.3
0.4	0.16	6.8	6.1
0.35	0.12	7.8	7
0.3	0.09	9	8
0.25	0.06	10.7	9.3
0.2	0.04	12.7	10.7
0.15	0.02	16.6	13
0.1	0.01	21.5	15.3

It appears that the datings obtained by (13) are much more reliable than those of classical glottochronology. For example, we date the disintegration of Belorussian and Ukrainian (97% correspondences) to be the fourteenth century A.D.; for different pairs of Germanic, Romance, Slavic or Turkic languages we get datings in the first millennium; for the disintegration of the Balto-Slavonic languages we get a dating around the end of the second millennium B.C. Of course, one should not overemphasize the precision of glottochronological data, since there may be various types of statistical fluctuation and perturbation<sup>24</sup>. However, it seems to be an

<sup>24</sup> Language contact is an obvious disturbing factor. This can produce secondary linguistic convergence of two already diverged languages. In this case often the fifth postulate of glottochronology is violated: in languages which are in active contact there is a tendency towards retention and/or loss of the same words from the BL. In such cases we cannot always speak of borrowing. Sometimes such contacts can cause an increase in the percentage of retentions by as much as 5 or 6 percent. We can observe this situation for Belorussian and West Slavic languages, for German and Scandinavian, etc. There are some ways of taking account of this effect while evaluating the degree of closeness, but all of these have a restricted application and they deserve separate consideration. An attempt of a general discussion of this problem was suggested by Hattori Shiro (1954) who suggested the constant 2 in the formula  $N_2(t) = e^{-2t}$  be replaced by a variable. This variable depends upon the degree of closeness between the languages. However it remains obscure how this coefficient is to be established.



important adjunct in genetic classification and in the evaluation of the closeness of related languages.

Everything said above relates to Swadesh's 'standard' one-hundred word BL. The existence of differences in individual retention rates suggests that, in principle, it is possible to compile lists which would differ from each other not only in the general coefficient  $\lambda$ , but also in their formulae for the correlation of rate of loss and time. It is not impossible that by varying the set of words in the list one could compile lists, which satisfy different correlation formulae. This work, however, would be computationally complex<sup>25</sup>.

Classical lexicostatistics, even with these improved dating methods, is still plagued by major shortcomings. There are well-known problems connected with the choice of the main word where two or more synonyms exist - for example, what word should be chosen in Italian for 'head' - *capo* or *testa* ? - or 'sand' - *sabbia* or *rena* ? - and so on. Apart from that, the lexicostatistical procedure lacks statistics in the strict sense: in comparing the BLs of two languages, we get but a single result, while to increase the reliability of results one would like to have a series of outcomes, from which it would be possible to calculate the mathematical confidence and the limits of possible fluctuation.

---

<sup>25</sup> In an experiment we combined words from the 100 word list as well as from the 200 word list of M. Swadesh and tried to compile a 55 word list which was to fulfil the classical "radioactive" equation  $N(t) = Ne^{-\lambda t}$  where  $\lambda = 0.1$  (i.e. "where the value of the coefficient of persistence was 0.9 for 1,000 years). This list contained the following words: *bark, belly, big, black, blood, bone, to burn, to die, dog, dry, ear, to eat, egg, eye fire, foot, full, hair, head, I, knee, leaf, liver, long, many, meat, moon, near, night, nose, round, short, snake, star, stone, swim, tail, this, thin, thou, tongue, tree, two, water, we, what, white, woman, worm, year, yellow, who, neck, new, mouth*. We tried to compile this list in such a way that it satisfied the procedure of 'classical' glottochronology (in particular, borrowings were counted as replacements). This list, which has been tested on relatively large linguistic material, allows us to build classificatory schemes which are reliable on the whole, and to find datings which match the datings according to the standard list of Swadesh (although with sometimes quite considerable deviations because of the smaller number of words); for short historical time spans though, datings still are definitely too 'young'.

The list of 35 most stable words with  $\lambda = 0.07$  or  $0.08$  has been compiled by Jaxontov. This list is very useful for verification of genetic relationship between languages. The percent of cognates among these 35 words should be higher than the percent between the remaining words of the list. However this short list is not quite suitable for dating or classification.

I assume that both these shortcomings could be eliminated with the help of a method which I call 'etymological statistics', or 'root glottochronology'. Let us formulate its main postulates.

1. In every language there are some roots that are original, i.e. not borrowed during the period of separate existence of this language. According to preliminary estimates, there are not much more than two or three thousand roots of this type in any modern language.

2. These roots have different frequencies of occurrence, in other words they have different probabilities of being found in a chosen text.

3. The frequency of occurrence (as just defined) of a given root in some language at a fixed period of time  $t$  is stable, and does not depend (or hardly depends) on the type of text.

4. All roots can 'age' - their frequency of occurrence gradually approaches zero, after which the root is considered to have disappeared from the language. At the same time, however, the rate of loss of different roots is not identical: roots, like words, may be divided into stable and less stable.

5. The loss of roots from a language proceeds at a steady rate - that is, from some set of roots, characterized by a fixed frequency, over a given period  $\Delta t$ , a fixed number of roots will be lost.

For this theory, the key postulate is 3, and a priori it raises most doubts. Indeed, the nature of the lexicon and the distribution of word frequency in texts of various genres vary considerably, except for such extremely frequent words as pronouns and grammatical morphemes. But one etymological root usually serves as a source for many derived words with different meanings. Since the frequency of a root correlates strongly with its productivity, the most frequent roots are usually found in stems belonging to very different semantic domains. Thus the higher the productivity and frequency of a root the better the chance of finding it in texts of any genre and subject matter. In any case, the empirical data clearly demonstrate the genre-independent and neutral character of the set of roots found in any text.

The fifth postulate of 'root glottochronology' is equivalent to the third postulate

of standard lexicostatistics (see above), and requires empirical confirmation (see below).

Let us take some samples - for example one hundred non-loaned roots from a language A - and let us supply each of these roots with their etymological correspondents in related languages. (I should note that this procedure necessarily presupposes prior etymological analysis). It is obvious that in those languages that are most closely related to language A, one can find the highest number of correspondences, while with increasing genetic distance the number of such correspondences will diminish. It is natural to expect that Russian, for example, will have more common roots with Slavic languages than with Baltic, and with the latter - more than with all other Indo-European languages, etc. This procedure would allow us to create classifications and relative chronologies of divergence within a language family. The analysis of several such samples should in principle give comparable results.

However, due to postulates (2) and (4), the absolute figures of etymological correspondences among roots of language A and related languages will vary considerably in random samples. In order to make these figures stable, one should select samples characterized by one and the same distribution of root frequencies.

One could compile dictionaries of root frequency and take one's samples from them. This would mean extending to the study of roots the work carried out for words by Arapov and Herz. However, such work is difficult to accomplish.

Here postulate (3) comes into play, according to which each root in the language has a given probability of being found in any text. Accordingly, any text should exhibit the same or similar distribution of frequencies of the roots represented in it.

If this is so, one would expect that in samples of genuine roots in different texts from language A there would be the same or a similar number of etymological correspondences with each of the related languages.

Let us take some English text - for example, the text of this article. We will choose from it non-loaned lexical roots; all prefixes, suffixes and proper names will be excluded, and each root will be counted only once. For each morpheme of this type

we shall look for etymological correspondences in German, Russian, Lithuanian, and French. We will not count cases where etymological correspondences are present in these languages, but only as loans.

English	German	Russian	Lithuanian	French	IE
1. the, this, that	+ (der)	+ (ТОТ)	+ (ta-s)	+ (te-l)	*to-
2. last	+ (letzt)	+ (лень)	+ (léid-)	+ (las)	*lē(i)-
3. two	+ (zwei)	+ (два)	+ (dù)	+ (deux)	*duō-
4. have	+ (haben)	-	-	+ (re-cev-)	*kap-
5. wit-ness	+ (wissen)	+ (вед-)	+ (vīd-)	+ (voir)	*weid-
6. a, once, any	+ (ein)	+ (од-ин)	+ (vīenas)	+ (un)	*oino-
7. in	+ (in)	+ (в)	+ (ĩ)	+ (en)	*en-
8. of	+ (ab)	+ (по)	+ (pa-, po)	+ (po-ndre)	*apo/*po
9. work	+ (wirken)	-	-	-	*werg-
10. with	+ (wider)	+ (второй)	-	-	*wi-

11. wide	+	-	-	-	(*weit-)
	(weit)				
12. laid	+	+	-	+	*leg-
	(legen)	(лежать)		(lit)	
13. for, from	+	+	+	+	*per, *pro-
	(für)	(про, при)	(pro)	(pour)	
14. or, other,	+	+	+	-	*eno-, *no-
be-yond, and	(ander-,jener)	(он)	(ana-s)		
15. step	+	-	-	-	*steb-
	(Stapfe)				
16. was	+	-	-	-	*wes-
	(war)				
17. by	+	+	-	+	*(o)bhi
	(bei)	(o[b])		(oub-lier)	
18. his, hinder	+	+	+	+	*'ke-, *'ki-
	(hier, hinter)	(з-десь)	(šis)	(ce)	
19. to-wards	+	+	+	+	*wert-
	(werden)	(вертеть)	(versti)	(vers)	
20. which,	+	+	+	+	*k <sup>w</sup> o-
who	(wer, welcher)	(к-то, ч-то)	(ka-s)	(que)	
21. begin	+	-	-	-	(*ghen-?)
	(beginnen)				
22. to	+	+	+	+	*do-/*de-
	(zu)	(до)	(-da)	(de)	
23. at	-	-	-	+	*ad
				(a)	

24. old	+	-	-	+	*al-
	(alt)			(haut)	
25. world	+	-	-	+	*wīro-
	(Welt)			(vertu)	
26. new	+	+	+	+	*newo-
	(neu)	(новый)	(naujas)	(neuf)	
27. if, it	+	+	+	+	*e-, *i-
	(ob, es)	(э-тот)	(jī-s)	(ce < ecce)	
28. we	+	-	+	-	*we-
	(wir)		(vè-du)		
29. as, also,	+	-	+	-	*al-
all	(als, all)		(al-vienas)		
30. sixth	+	+	+	+	*s(w)ek's-
	(sechs)	(шесть)	(šeši)	(six)	
31. through	+	-	-	+	*ter-
	(durch)			(tres, tra-)	
32. out, b-ut,	+	+	+	+	*ud-
a-b-out	(aus)	(вы-)	(už-)	(j-usque)	
33. is	+	+	+	+	*es-
	(ist)	(есть)	(es-)	(es-t)	
34. are	-	-	+	-	*er-/*or-
			(yга)		
35. arise	+	+	+	+	* (o)rei-
	(reisen)	(пой)	(ry'-tas)	(ruisseau)	
36. so, such	+	+	+	+	*s(w)e
	(so, sich)	(свой, себя)	(sāvo)	(se)	

37. can, know	+	+	+	+	*gnō-
	(kann)	(знать)	(žino <sup>́</sup> -ti)	(connaître)	
38. must	+	-	-	+	*med-
	(muß)			(co-mme)	
39. now	+	+	+	-	*nũ-
	(nun)	(ныне)	(nu <sup>`</sup> )		
40. answer	+	+	-	-	*swer-
	(schwören)	(свара)			
41. be	+	+	+	+	*bheu-
	(bi-n)	(быть)	(bũ-ti)	(fu-t)	
42. same,	+	+	+	+	*sem-
some	(zu-sammen)	(сам)	(san-)	(sim-ple)	
43. rather	+	-	+	-	*kret-
	(retten)		(krēs-ti)		
44. would,	+	+	+	+	*wel-
well	(wollen)	(воля)	(valia <sup>`</sup> )	(vouloir)	
45. not	+	+	+	+	*ne
	(nicht)	(не)	(ne <sup>`</sup> )	(non)	
46. our, us	+	+	-	+	*no-(s)
	(uns)	(нас)		(nous)	
47. kind	+	-	-	+	*gēnə-
	(Kind)			(naître)	
48. least	-	-	-	-	(*leis-)
49. should	+	-	+	-	*skel-
	(sollen)		(skelē <sup>́</sup> -ti)		
50. earlier	+	-	-	-	*ajer-
	(eher, erst)				

51. much	-	-	-	+	*mag- (maire, mais) (*meg)
52. more, most	+	-	-	-	*mē-
	(mehr, meist)				
53. on	+	+	+	-	*an-a, *an-ō
	(an)	(на)	(nuō)		
54. ever, every	+	-	-	+	*aiw- (âge)
	(ewig)				
55. ten	+	+	+	+	*dek-m- (dix)
	(zehn)	(десять)	(dēsimt)		
56. thousand	+	+	+	-	*tūs-
	(Tausend)	(тысяча)	(tūkstantis)		
57. year	+	+	-	-	*iōr-
	(Jahr)	(яровой)			
58. time	+	-	-	-	*dā(i)-
	(Zeit)				
59. depth	+	+	+	-	*dheub-
	(tief)	(дно)	(dubus)		
60. in-deed, do	+	+	+	+	*dhē-
	(tu-n)	(де-ть)	(dē-ti)	(fai-re)	
61. small	+	+	-	+	*(s)māl-
	(schmal)	(малый)		(mal, mauvais)	
62. greater	+	+	+	-	*ghreud-
	(groß)	(груда)	(grūs-ti)		
63. like-wise	+	-	+	-	*līg-
	(gleich)		(lýgus)		



64. deal	+	+	+	-	*dhoi-l-
	(Teil)	(делить)	(daily'-ti)		
65. call	+	+	-	-	*gol(-s-)
	(klagen)	(голос)			
66. tree	+	+	+	-	*derw-
	(treu, Teer)	(дерево)	(derva')		
67. wood	-	-	-	-	*widhu-
68. few	-	-	-	+	*pau-
				(pauvre)	
69. find	+	+	-	+	*penth-
	(finden)	(путь)		(pont)	
70. root	+	-	-	+	*wərad-
	(Wurzel)			(racine)	
71. side, since	+	-	-	+	*sēj-
	(Seite, seit)			(po-nd-re)	
72. back	-	-	-	-	(*bhag-)
73. may	+	+	+	-	*megh-
	(mag)	(мочь)	(mēg'-ti)		
74. stem	+	-	+	-	*ste(m)bh-
	(Stamm)		(stemb'-ti)		
75. set, settle	+	+	+	+	*sed-
	(sitzen)	(сидеть)	(sēdē'-ti)	(as-seoir)	
76. even	+	-	-	-	(*ebh-?)
	(eben)				
77. still	+	+	+	+	*stel-
	(still)	(стол, столб)	(stālas, stuļbas)	(lieu)	

78. East	+	-	+	-	*awes-
	(Osten)		(aušrà)		
79. West	+	-	-	+	*wes-
	(Westen)			(vêpres)	
80. hand	+	-	-	-	(*kon- ~ *k-, -m-)
	(Hand)				
81. ninth	+	+	+	+	*new-m
	(neun)	(девять)	(devyni)	(neuf)	
82. ask	+	+	+	-	*ais-
	(heischen)	(искать)	(ieško'ti)		
83. over,	+	+	-	+	*upo
often	(über)	(высокий)		(sous)	
84. give	+	-	+	+	*ghabh-
	(geben)		(gabénti)	(avoir)	
85. young	+	+	+	+	*jewən-
	(jung)	(юный)	(jaunas)	(jeune)	
86. be-long	+	+	+	+	*delə(n)gh-
	(lang, gelangen)	(долгий)	(ilgas)	(long)	
87. many	+	+	+	-	*menegh-
	(manch)	(много)	(minia)		
88. higher	+	+	+	-	*keuk-
	(hoch)	(куча)	(kaũkaras)		
89. show	+	+	-	-	*(s)kew-
	(schauen)	(чуять)			
90. fully	+	+	+	+	*pelə-
	(voll)	(полный)	(pilnas)	(plein)	

91. I	+	+	+	+	*eg(h)om
	(ich)	(я)	(aš)	(je)	
92. think	+	-	-	-	*tong-
	(denken)				
93. word	+	+	+	+	*wer-
	(Wort)	(врать)	(vardas)	(verve)	
94. half	+	-	-	-	(*kalp-)
	(halb)				
95. a-go	+	-	-	-	*ghē-
	(gehen)				
96. share	+	-	+	-	*(s)ker-
	(scheren)		(skir-ti)		
97. five	+	+	+	+	*penk <sup>w</sup> e
	(fünf)	(пять)	(penki)	(cinq)	
98. start	+	+	+	+	*ster-
	(stürzen)	(страдать)	(starinti)	(étrene)	
99. talk	+	-	-	+	*del-
	(Zahl)			(deuil)	
100. say	+	-	-	-	*sagh-
	(sagen)				

Thus in this sample of text among 100 non-loaned roots<sup>26</sup> we find (if we score doubtful etymologies at 0.5%) 94 roots shared with German, 58.5 with Lithuanian, 60 with Russian, and 58 with French.

<sup>26</sup> From 100 selected morphemes, *begin*, *back*, *even*, *hand* and *half* do not have IE etymologies. Nevertheless it is not proven that these morphemes have been borrowed. Thus we have retained them in the list. In any case, there are too few of them to have any significant influence on the results of our analysis. On the other hand, we excluded from the list many stems whose status as borrowings is well-proven. An average English text contains in fact many more borrowings than inherited roots which may be actually the reason for a slight increase of the cognacy rates, compared with respective figures for Russian, German and other texts.

It turns out that these figures are quite stable. Let us compare the results of the calculations done with some Russian texts (with the text number in brackets - see page below):

	Polish	Lithuanian	German	French
(6)	98	77	55	52
(10)	96	76	58	50
(4)	97	70	51.5	50
(11)	93	72	55	51.5

We see that the figures in each case cluster around some statistical mean: for Polish it is  $95 \pm 3\%$ , for Lithuanian  $74 \pm 3\%$ , for German  $54 \pm 3\%$ , and for French  $51 \pm 1\%$ .

We shall now give some more results for different pairs of languages. By 'text language' we mean the language whose text is being analyzed, and by 'dictionary language' the language whose dictionary is used for the comparison.

Text Language	Dictionary Language	Proportion of Correspondences	Text
French	Russian	0.50	5
German	Russian	0.55	2
German	Lithuanian	0.57	2
Latin	Russian	0.50	3
Latin	German	0.55	3
Latin	Lithuanian	0.53	3
Ancient Greek	Vedic	0.69	9
Ancient Greek	Latin	0.67	9
Ancient Greek	Russian	0.52	9
Ancient Greek	German	0.54	9
Ancient Greek	Lithuanian	0.53	9
Latin	Vedic	0.71	3

Latin	Ancient Greek	0.72	3
Russian	Lithuanian	0.76	7
Russian	Lithuanian	0.74	8
Vedic	Ancient Greek	0.77	1
Vedic	Latin	0.64	1
Vedic	Russian	0.54	1
Vedic	Lithuanian	0.53	1
Vedic	German	0.57	1

Besides the relatively high stability of correspondence levels, we observe an interesting phenomenon. While calculating from 'text language' to 'dictionary language' and vice versa (i.e. when the roles of 'text language' and 'dictionary language' are exchanged) the figures for languages of the same period are similar. That the figures should increase, when comparing modern 'text languages' with ancient 'dictionary languages' is only to be expected. But the calculations from ancient 'text languages' to modern 'dictionary languages' reveal practically the same figures as when we compare two modern languages. One can thus formulate a very important rule: the age of the text does not influence the result of the statistical analysis.

With closer consideration this result becomes quite understandable. The conditions required by the method in each case let us measure not the distance from one language to another, but the distance from the proto-language to the dictionary language<sup>27</sup>, and therefore we obtain only the figures characterizing this distance.

Let us try one more tack. We shall take as a sample text the one-hundred word Swadesh list and will apply the same procedure to it. That is, we will write out all non-loaned roots, ignoring borrowings and repetitions. Let us analyze this material with the help of the same languages as above, comparing German, Russian, Lithuanian and French with English.

---

<sup>27</sup> Translator's Note: Since we are in fact measuring how many roots reflected in the sample are retained in the dictionary language.- NE & IP.

	English	German	Russian	Lithuanian	French	IE
1.	all	† (als, all)	-	† (al-vienas)	-	*al-
2.	ashes	† (Asche)	-	-	† (ardeur)	*as-
3.	belly	† (Balg)	-	-	-	*bhelǵh-
4.	bird	-	-	-	-	(*bhreu-)?
5.	bite	† (beißen)	-	-	† (fendre)	*bheid-
6.	black	† (blecken)	-	-	† (foudre)	*bhelg-
7.	blood	† (Blut)	-	-	† (fleur)	*bhele-
8.	bone	† (Bein)	-	-	-	(*bhoin-)
9.	breast	† (Brust)	† (брюхо)	-	-	(*bhreus-)
10.	burn	† (brennen)	† (бруить)	† (briautis)	-	*bhreu-
11.	cloud	-	-	-	-	*gleut-
12.	cold	† (kalt)	†? (холод)	†? (šáltas)	† (gel)	*gel-

13.	come	+	-	+?	+	*g <sup>w</sup> em-
		(kommen)		(gimti)	(venir)	
14.	die	+	-	-	-	*dheu-
		(tot)				
15.	drink	+	-	+	-	*dhreg-
		(trinken)		(drēž-ti)		
16.	dry	+	-	-	-	*dhreugh-
		(trocken)				
17.	ear	+	+	+	+	*ous-
		(Ohr)	(yxo)	(ausis)	(oreille)	
18.	earth	+	-	-	-	*er(t)-
		(Erde)				
19.	eat	+	+	+	-	*ed-
		(essen)	(есть)	(ēsti)		
20.	eye	+	+	+	+	*ok <sup>w</sup> -
		(Auge)	(око)	(aki-s)	(oeil)	
21.	fat	+	+	+	-	*poi-
		(feist)	(питать)	(pydyti)		
22.	feather	+	+?	+	+	*pet-/
		(Feder)	(перо)	(sparnas)	(pennē) *pter- <sup>28</sup>	
23.	fire	+	-	-	-	*peHwōr
		(Feuer)				
24.	fish	+	+?	-	+	*peisk-
		(Fisch)	(пескаръ)		(poisson)	

<sup>28</sup> As we have Greek πτερόν it seems reasonable that the IE word for "feather" was originally connected with IE \*pet- "to fly". Due to development from \*pter- to \*per- in different languages including Balto-Slavic, there were opportunities of contamination with the IE root \*per- "to move", Russian *parit'* and German *fahren* etc.

25.	fly	†	†	†	†	*pleu-
		(fliegen)	(плыть)	(pláuti)	(pluie)	
26.	foot	+	†	†	†	*ped-
		(Fuß)	(под)	(pādas)	(pied)	
27.	full	+	+	+	+	*pelə(n)-
		(voll)	(полный)	(pilnas)	(plein)	
28.	give	+	-	†	†	*ghabh-
		(geben)		(gabēnti)	(avoir)	
29.	go	+	-	-	-	*ghē-
		(gehen)				
30.	good	+	+	+	-	*ghadh-
		(gut)	(годный)	(guōdas)		
31.	green	+	-	-	-	*ghrē-
		(grün)				
32.	hair	+	+	+	-	*ker(s)-
		(Haar)	(шерсть)	(šerys)		
33.	hand	+	-	-	-	(*kon- ~
		(Hand)				*k-, -m-)
34.	head	+	-	-	+	*kap-
		(Haupt)				
35.	hear	†	+	-	-	*(s)kew-
		(hören)	(чуять)			
36.	heart	+	+	+	+	*kerd-
		(Herz)	(сердце)	(širdis)	(coeur)	
37.	horn	+	+	+	+	*kern-
		(Horn)	(корова)	(karve)	(corne)	



38.	I	+	+	+	+	*eghom
		(ich)	(я)	(aš)	(je)	
39.	kill	+	+	+	-	*g <sup>w</sup> el-
		(quālen)	(жалъ)	(gēlti)		
40.	knee	+	+	-	+	*g <sup>u</sup> enu-
		(Knie)	(звено)		(genoux)	
41.	know	+	+	+	+	*gnō-
		(kennen)	(знать)	(žinoti)	(connaitre)	
42.	leaf	+	+	+	+	*leubh-
		(Laub)	(луб)	(lubà)	(livre)	
43.	lie	+	+	-	+	*legh-
		(liegen)	(лежать)		(lit)	
44.	liver	+	-	-	-	*leprò-
		(Leber)				
45.	long	+	+	+	+	*delə(n)gh-
		(lang)	(долгий)	(ilgas)	(long)	
46.	louse	+	+?	+?	-	*wes- ~
		(Laus)	(вошь)	(vievisa)		*lūs-
47.	man	+	+	-	-	*mon-
		(Mann)	(муж)			
48.	many	+	+	+?	-	*menegh-
		(manch)	(много)	(minia)		
49.	meat	+	-	-	-	*mad-
		(Mast)				
50.	moon	+	+	+	+	*mēnes-
		(Mond)	(месяц)	(mēnuo)	(mois)	

51.	mouth	+	-	-	† *men-
		(Mund)			(menton)
52.	nail	+	+	+	†(o)nogh <sup>w</sup> -
		(Nagel)	(нога)	(naga, nāgas)	(ongle)
53.	name	+	+	-	† *(e)nomn-
		(Name)	(имя)		(nom)
54.	neck	+	-	-	-
		(Nacken)			
55.	new	+	+	+	† *new-
		(neu)	(новый)	(naũjas)	(neuf)
56.	night	+	+	+	† *nok <sup>w</sup> t-
		(Nacht)	(ночь)	(naktis)	(nuit)
57.	nose	+	+	+	† *nas-
		(Nase)	(нос)	(nosis)	(nez)
58.	not	+	+	+	† *ne-
		(nicht)	(не)	(nè)	(non)
59.	one	+	+	+	† *oj-n-
		(ein)	(один)	(vienas)	(un)
60.	rain	+	-	+	- *rek-
		(Regen)		(rōk-ti)	
61.	red	+	+	+	† *reudh-
		(rot)	(рыжий)	(raũdas)	(rouge)
62.	road	+	-	-	- *reidh-
		(reiten)			
63.	root	+	-	-	† *wərad-
		(Wurzel)			(racine)

64.	sand	+	-	-	-*sandh-(?)
	(Sand)				
65.	say, see	+	-	†	- *sek <sup>w</sup> -
	(sagen, sehen)			(sakyti)	
66.	seed	+	†	†	+*sē-(men)-
	(Same)	(семя)	(sēmens)	(semence)	
67.	sit	+	+	+	+ *sed-
	(sitzen)	(сидеть)	(sēdėti)	([as]-sis)	
68.	sleep	†	+	+	- *sleb-
	(schlafen)	(слабый)	(slābnas)		
69.	small	+	+	-	+ *(s)mal-
	(schmal)	(малый)		(mal)	
70.	smoke	+	+	+	- *smeug-
	(schmauchen)	(смуглый)	(smaugti)		
71.	stand	+	+	+	+ *stā-
	(stehen)	(стоять)	(stó-ti)	(être)	
72.	star	+	-	-	+ *Haster-
	(Stern)			(étoile)	
73.	stone	+	+	-	- *stei-
	(Stein)	(стена)			
74.	sun	+	+	+	+ *swel-
	(Sonne)	(солнце)	(saulē)	(soleil)	
75.	swim	+	-	+	- *swem-
	(schwimmen)		(sumdyti)		
76.	tail	-	-	-	- *dēk-
77.	that,	+	+	+	+ *to
	this	(der)	(тот)	(ta-s)	(tel)

78.	tongue	+	†	†	+ *ienghu- /
		(Zunge)	(язык)	(liežūvis)	(languc)*denghu-
79.	tooth	+	-	†	+ *dent-
		(Zahn)		(dantis)	(dent)
80.	tree	+	+	+	- *derw-
		(treu, Teer)	(дерево)	(derva)	
81.	two	+	+	†	+ *dwo(u)
		(zwei)	(два)	(du)	(deux)
82.	warm	+	+	+	+ *g <sup>w</sup> her-
		(warm)	(гореть)	(gāras)	(four)
83.	water	+	+	+	- *wed-
		(Wasser)	(вода)	(vanduō)	
84.	we	+	-	+	- *we-
		(wir)		(vè-du)	
85.	who,	+	+	+	+ *k <sup>w</sup> e-/
	what	(wer, was)	(кто, что)	(ka-s)	(que) *k <sup>w</sup> i-
86.	white	+	+	+	+ *k <sup>w</sup> ei-
		(weiß)	(свет)	(šviesti)	(verre)
87.	woman	+	-	-	- *weip-?
		(Weib)			
88.	yellow	+	+	+	+ *gh <sup>w</sup> el-
		(gelb)	(желтый)	(gel̃tas)	(fiel)
89.	you	+	-	+	- *ju-
		(ihr)		(jūs)	

Out of 89 roots from the English one-hundred word list<sup>29</sup> we have 87

<sup>29</sup> Eight loans, *bark*, *big*, *dog*, *egg*, *mountain*, *person*, *round* and *skin* have been eliminated from the list. Pairs of words with similar roots are combined. These include *say*, *see*, *who*, *what*; *that*, *this*.

correspondences with German (98%); 52 correspondences with Russian (58%); 51 correspondences with Lithuanian (57%) and 48 correspondences with French (54%).

Thus we obtain practically the same results as were obtained in our discussion of samples of English and Russian roots from arbitrarily chosen texts. This evidently shows that the distribution of individual frequency characteristics of roots in the Swadesh list coincides with their usual distribution in texts<sup>30</sup>. From this one can reach some important conclusions:

(a) the 'stability' of roots does not depend on the 'stability' of the words derived from them. This follows first of all from the third postulate of root glottochronology, but is also well demonstrated by the example of the Swadesh list discussed above: words included in it are intentionally more 'stable' than most words in a randomly chosen text, while the stability of roots is the same.

(b) in the absence of texts (which unfortunately is often the case with many lesser-known languages) the Swadesh one-hundred word list could be used as a text for root glottochronology.

(c) the mathematical apparatus worked out for classical glottochronology could be transferred to root glottochronology, but obviously requires another value for  $\lambda$ .

The third conclusion is the strongest, and will acquire additional practical experimentation. According to the preliminary results, however, the time calculated by inserting the proportion of root correspondences into the formula

$$(14)^{31} \quad t = \sqrt{\frac{\ln(N(t)/N_0)}{-\lambda N(t)}} \quad \text{with } \lambda = 0.035$$

corresponds in general with the datings obtained by the version of classical glottochronology proposed above with formula (13). Compare the dates of divergence

<sup>30</sup> The explanation of this is perhaps the fact that the Swadesh list includes the most common and the most usual ideas from quite different semantic fields, which thus really creates some kind of elementary conceptual text concerning humans and their environment. It is remarkable that many (22!) English roots from the Swadesh list are represented in the text discussed above: full, what/ who, go, root, long, that/this, new, we, know, give, one, many, all, two, lie, not, small, tree, sit, hand, hear/show, I.

<sup>31</sup> Recently D. Leshshiner has suggested a different formula for dating language divergence within root glottochronology:

of Russian from Polish, Lithuanian, German, and French (with time given in millennia)

	Polish <sup>32</sup>	Lithuanian <sup>33</sup>	German <sup>34</sup>	French <sup>35</sup>
t according to classical glottochronology	1.3	3.1	4.7	4.7
t according to root glottochronology	1.2	3.2	4.9	5.1

$$t = \frac{1 - \ln(N(t))}{1 - \ln(1 - N(t))} * T$$

where T (the period of "halfdecay") is equal to 5.

<sup>32</sup> In Russian and Polish the following words from the 100 word list do not match: *život* - *brzuch*, *bol'soj* - *wielki* (*duży*), *grud'* - *piers*, *žeč'* - *palić*, *glaz* - *oko*, *žir* - *tuszczy*, *xorošij* - *dobry*, *pečen'* - *wątroba*, *mnogo* - *wiele* (*dużo*), *luna* - *księżyc*, *rot* - *usta*, *krasnij* - *czerwony*, *skazat'* - *powiedzieć* (*rzec*), *koža* - *skóra*, *xvost* - *ogon*, *ženščina* - *kobieta*, *xolodnyj* - *zimny*. Taking two borrowings in the Russian list into account (*oblako* and *sobaka*) we arrive at 85 % matches.

<sup>33</sup> In Russian and Lithuanian the following words from the 100 word list match: *ves'* - *višas*, *kusat'* - *kąsti*, *krov'* - *kraujas*, *žeč'* - *dėgti*, *nogot'* - *nėgas*, *xolodnyj* - *šaltas*, *pri-jti* - *ateiti*, *umeret'* - *mirti*, *suxoj* - *sausas*, *uxo* - *ausis*, *zemlja* - *žemė*, *ogon'* - *ugnis*, *letet'* - *lėkti*, *polnyj* - *pilnas*, *dat'* - *duoti*, *zelenyj* - *žalias*, *ruka* - *ranka*, *golova* - *galvė*, *serdce* - *širdis*, *rog* - *rągas*, *ja* - *ąš*, *znat'* - *žinoti*, *koleno* - *kelis*, *pečen'* - *kepenys*, *dlinnyj* - *ilgas*, *mjaso* - *mėsà*, *novyj* - *naujas*, *nač'* - *naktis*, *nos* - *nosis*, *ne* - *nė*, *odin* - *vienas*, *sem'a* - *sėda*, *sidet'* - *sėdėti*, *dym* - *dūmai*, *stoyat'* - *stovėti*, *zvezda* - *žvaigždė*, *kamen'* - *akmuo*, *solnce* - *saulė*, *plavat'* - *plaukti*, *tot* - *tàs*, *ty* - *tù*, *jazyk* - *liežuvis*, *dva* - *dù*, *voda* - *vanduo*, *my* - *mes*, *č-to* - *kàs*, *belyj* - *baltas*, *želnyj* - *geltonas*; thus (not taking the words *oblako* and *sobaka* into account) exactly 50 % matches.

<sup>34</sup> In Russian and German the following words from the 100 word list form matches: *nogot'* - *Nagel*, *uxo* - *Ohr*, *est'* - *essen*, *jajco* - *Ei*, *pero* - *Feder*, *polnyj* - *voll*, *serdce* - *Herz*, *ja* - *ich*, *znat'* - *kennen*, *ležat'* - *liegen*, *dlinnyj* - *lang*, *mužčina* - *Mann*, *imja* - *Name*, *novyj* - *neu*, *noč'* - *Nacht*, *nos* - *Nase*, *ne* - *nicht*, *odin* - *ein*, *sem'a* - *Same*, *sidet'* - *sitzen*, *stoyat'* - *stehen*, *solnce* - *Sonne*, *etot* - *dieser*, *ty* - *du*, *jazyk* - *Zunge*, *dva* - *zwei*, *voda* - *Wasser*, *č-to* - *was*, *kto* - *wer*. Taking into account two borrowings in the Russian list (*oblako*, *sobaka*) and three borrowings in the German list (*Fett*, *Kopf*, *rund*) we arrive at 30% matches.

<sup>35</sup> In Russian and French the following words from the 100 word list match: *kora* - *écorce*, *nogot'* - *ongle*, *umirat'* - *mourir*, *pit'* - *boire*, *uxo* - *oreille*, *jajco* - *oeuf*, *polnyj* - *plein*, *dat'* - *donner*, *serdce* - *coeur*, *ja* - *je*, *znat'* - *connaître*, *dlinnyj* - *long*, *luna* - *lune*, *imja* - *nom*, *novyj* - *neuf*, *noč'* - *nuit*, *nos* - *nez*, *ne* - *ne*, *odin* - *un*, *videt'* - *voir*, *sides'* - (*être*) *assis*, *dym* - *fumée*, *stojat'* - (*être*) *debout*, *solnce* - *soleil*, *ty* - *tu*, *jazyk* - *langue*, *dva* - *deux*, *čto* - *que*, *kto* - *qui*. Taking into account two borrowings in the Russian list (*oblako*, *sobaka*) and three borrowings in the French list (*chemin*, *pietre*, *blanc*) we arrive again at 30 % matches.

Some of these datings are possibly too early. This is particularly true for the Russian-Polish divergence, where the increase in the percentage of correspondences could be due to secondary contacts: other Slavic languages allow us to date the Slavic divergence 200 to 300 years earlier. But in general they seem quite reasonable.

The method of 'root glottochronology' has some advantages over classical glottochronology. In particular, there are no problems with the choice of a 'main' word from several synonyms. It is also possible, to carry out a statistical analysis on a number of results, because the number of texts sampled - in contrast to the BL list - is in principle unlimited<sup>36</sup>. We can foresee, however, that etymostatistical analysis will face major obstacles when applied to languages whose history is less known. I would underline, however, that a thorough comparative historical analysis of the language data should precede any lexicostatistical or etymostatistical study, which will in any case be complementary to rather than a substitute for, comparative work.

Etymostatistics is currently being applied successfully to Semitic and Afro-Asiatic language materials by A. Militarev, and to Austronesian by Y. Sirk.

As we can see, the combination of lexicostatistics and etymostatistics allows us to obtain more precise datings and classifications, both for normal families and macro-families. According to preliminary results, modern Nostratic languages reveal levels of 15-20% correspondence according to root glottochronology, when one compares randomly-chosen texts with reliable dictionaries. These are much higher percentage values than those found between modern Nostratic languages on the 100-word Swadesh list (on average 5-9%), and I think that in the future, when we have a better understanding of comparative phonology and etymology, root glottochronology will play an important role in testing theories of remote relationships and in creating genetic classifications.

-Translated by N. Evans and I. Peiros.

---

<sup>36</sup> One further possible application of glottochronology is the dating of texts. This is however a separate area, which has its specific problems, and we will not discuss them here.

TEXTS USED FOR ETYMOLOGICAL STATISTICS

- (1) Aufrecht T. *Die Hymnen des Rigveda, I.*  
Berlin 1995, pp. 1-2.
- (2) Böll H. *Wanderer, Konunst du nach Spa... (Erzählungen).*  
München 1971, pp. 7-8.
- (3) Caesar G. Julius. *Commentarii de bello Gallico (liber primus).*  
Moskva 1946, pp. 37-48.
- (4) Čukovsky K. *Stixi i Skazki.*  
Moskva 1984, pp. 37-39.
- (5) Dauzat A. *Dictionnaire étymologique de la langue française.*  
Paris 1938, pp. V-VI
- (6) Freidenberg O.M. *Mifi i literatura drevnosti.*  
Moskva 1978, pp. 206-207.
- (7) *Kulturnoe nasledie Vostoka (Problemy, poiski, suždenija).*  
Leningrad 1985, pp. 34.
- (8) Moloxovec E. - *Podarok molodym xoz'ajkam.*  
Sankt-Peterburg 1904, p. 288.
- (9) *Sophoclis tragoediae.*  
Moskva 1884, pp. 111-112.
- (10) Zošsenko M. *Izbrannoje v dvux tomax*  
Leningrad 1982, Vo. I, pp. 31-32.
- (11) Russian original of this paper,  
Moskva 1989.



LITERATURE

- Arapov, M., Herz, M. 1974: *Matematičeskie metody v istoričeskoj linguistike*.  
Moskva.
- Bergsland, K., Vogt, H. 1962: *On the validity of glottochronology*.  
In Current Anthropology, v. 3.
- Chrétien, D. 1952: *The Mathematical Model of Glottochronology*.  
In Language, v. 38.
- Dyen, J., James, A. 1967: *English Divergence and Estimated Word Retention Rate*.  
In Language, v. 47.
- Embleton, S. 1986: *Statistics in Historical Linguistics*.  
Bochum
- Fodor, J. 1961: *A glottochronologia ervenyessege a szlav nyelvek anyaga alapjan*.  
In Nyelvtudományi Közlemények, v. 63, No. 2.
- Guryčeva, M. 1959: *Narodnaya latyn'*  
Moskva
- Hattori Shiro: 1954: "*Gengo nendaiku*" sunawachi "*goito:keigaku*" no ho:ho: ni tsuite (*Nihon sogo no nendai*).  
In Hattori Shiro: Gengogaku no ho:ho:, Tokyo 1960, pp.515-566.
- Hattori Shiro: 1959: *Nihongo no keito*:  
Tokyo .
- Helimski, E. 1984: *Problemy granic nostratičeskoj makrosem'ji jazykov*.  
In Problemy izučenija nostratičeskoj makrosem'ji jazykov. Moskva, pp. 31-48.
- Helimski, E. 1986: *K izučeniju i nadežnosti indoevropejsko-semitskix leksičeskix sootvetstvij*.  
In Balkany v kontekste Sredizemnomor'ja. Moskva.
- IlliF-SvityF, V. OSTJa 1971: *Opyt sravnenija nostratičeskix jazykov, t. 1*,  
Moskva.
- Jaxontov, S, 1984: *Glottoxronologija: Trudnosti i perspektivy*  
In Drevnejšaja jazykovaja situacija v Vostočnoj Azii, pp. 39-47.
- Merwe, N.van der, 1966: *New Mathematics for Glottochronology*.  
In Current Anthropology, v. 7.
- Muller, H. 1929: *Chronology of Vulgar Latin*.  
In Zeitschrift für Romanische Philologie, Beiheft 78.
- O'Neil, W. 1954: *Problems in the Lexicostatistic Time Depth of Modern Icelandic and Modern Faroese*.  
In General Linguistics, v. VI, No. 1.

Swadesh, M. 1960: *Leksikostatističeskoe datirovanie doistoričeskix etničeskikh kontaktov.*

In Novoe v lingvistike, vyp. 1. Moskva.

Swadesh, M. 1960a: *K voprosu o povyšēnii točnosti v leksikostatističeskom datirovanii.*

In Novoe v lingvistike, vyp. 1. Moskva.

Swadesh, M. 1965: *Lingvističeskie sv'azi Ameriki i Evrazii.*

In Etimologija 1964, Moskva

## GLOTTOCHRONOLOGY: DIFFICULTIES AND PERSPECTIVES\*

Sergei Jaxontov

Glottochronology is a technique for dating the separation of genetically related languages by the number of cognate words they share in common.

It is obvious that genetically related languages diverge with the flow of time. If so, then the degree of differentiation between the two related languages allows us to judge about the age of their separation. Glottochronology can be regarded as an attempt to give a concrete dimension to these general considerations.

Glottochronology begins with a number of statements about possible changes in the lexicon. It is assumed that:

1) there exists a number of concepts, which are tied in the least degree possible to the specific culture and are so elementary that their designations can be found in any language independent of its geographic, historic and social environment;

2) the changes undergone by the sub-lexicon representing such concepts can be only unconditional;

3) the probability of these changes is more or less constant for all languages.

All these postulates should be regarded as no more than likely or approximate; the same applies to the findings obtained through the application of the method discussed.

M. Swadesh, the founding father of the glottochronological technique, originally drew up a list consisting of 215 common concepts. Languages were sought which could be documented for 1,000 years or more; on the basis of these, it was decided that about 19% of the elements on the list were substituted in a thousand years. A word is considered as substituted if it stops being the most common designation of the

---

\* *Lingvističeskaja rekonstrukcija i drevnejšaja istorija Vostoka. Tezisy i doklady konferencii. Pt. 4. Drevnejšaja jazykovaja situacija v Vostočnoj Azii.* Moscow: Nauka, 1984, p.39-47.

respective concept. Such a word does not necessarily disappear from the language; it can be applied to other meanings or even retain the original meaning functioning as a comparatively rare synonym of the new word which replaces it (cf., for instance, Russian *oko* 'eye' replaced by the word *glaz*). The probability of word retention (i.e. non-substitution) in a thousand years was found to be 0.81. This coefficient was termed "the index of retention" ( $r$ ). The percentage of the words on the list, retained in  $t$  millennia, is  $r^t$ .

Actual  $r$  values fluctuate from language to language; however, the deviation from the average is relatively insignificant. Moreover, the longer the period, the closer the actual value will be to the average.

Soon after the introduction of the method, words on the list were shown to differ in their persistence, i.e., in the probability of retention. Since it is an average of individual retention rates, the index of retention changes depending on the specific words selected. For the sake of greater accuracy, M. Swadesh reduced the diagnostic list to 100 items, and then the index of retention turned to be 0.86.

If two dialects separate, each will have an independent development. The possibility for one language to retain a certain word in 1000 years being estimated as  $r$ , the probability of its retention in both languages will be  $r^2$ .

For any period of time, the probable percentage of cognates shared by two languages can be determined. Conversely, the time of divergence can be obtained for any percentage of cognates. The latter is calculated by means of the following equation:  $t = \frac{\log(C)}{2 \log(r)}$

where  $C$  is the fraction of cognates (number of cognates divided by the total number of elements on the diagnostic list); time  $t$  is expressed in millennia.

The formula type employed in glottochronology is applicable to any random processes. The  $r$  value, as was already shown, is obtained empirically.

The practice of glottochronology implies selecting equivalents for the terms of the diagnostic list in the two languages studied and subsequently determining the number of cognates. For equivalents, the most conventional designations of the

respective concepts should be selected. Words treated as cognates must have identical meaning and go back to one and the same proto-form. Words, known to have been borrowed from one language pair into the other or from a third source, are excluded from the cognate set. The percentage of cognates  $C$  can be translated into the divergence time  $t$  by means of the above given equation.

It is advisable to bear in mind that the  $r$  value was determined on the basis of the two concrete lists to which it primarily refers. Some studies make use of regional lists (drawn up for a specific language family or geographical area). The divergence time should be calculated on the basis of such lists only after the index of retention is determined. This can be done indirectly, too, by comparing the number of cognates established by the standard and regional lists respectively across several pairs of languages.

The reliability of glottochronological rates is defined by standard deviation (the pertinent formula is omitted from this paper). In two out of three cases, the deviation of true values from the mean value does not exceed one standard deviation. For the number of cognates obtained from the 100 item list, the standard deviation will generally be 4 or 5 (it will not exceed 5, in any case). Translated into time, this will mean 12-15% of the period calculated if this period does not exceed 1500 years. Thus, if two languages are shown to have 4000 years of divergence, the probability will be 2 out of 3 that they actually separated 3500-4500 years ago.

The error can be reduced by bringing more than two languages into comparison (for example, by comparing several Slavic and several Germanic languages, rather than just Russian and English). Another way to increase accuracy in dating would be to apply a number of different diagnostic lists and compare the respective data.

Glottochronology stands as part of comparative linguistics or as a supplement to it. It fully accepts all the postulates and conclusions of the parent science. Strictly speaking, glottochronology can be applied only to the languages shown to be genetically related by conventional methods. Cognate words are then defined by etymological dictionaries or, at least, are identified on the basis of the established phonetic correspondences. Accordingly, conventional comparative methods are

applied to eliminate borrowings.

This would be the ideal case. However, language families have been studied quite unevenly. Hence, glottochronology has frequently been applied to languages the history of which either has not been studied at all or has been studied with no conclusive results (as is the case of the Altaic family, which, despite a long tradition of comparative study, still remains a tentative grouping). For such languages, tentatively related words, rather than definite cognates, are counted. This certainly diminishes the reliability of the respective glottochronological scales; however, they are not altogether worthless. Incorrect correspondences affect calculation results; in particular, they cause a certain "rejuvenation" of the languages under comparison. Nevertheless, one cannot expect genetically related languages to show superabundant similarities which are not due to the common origin; such similarities would hardly comprise 10% of non-cognate items. The same refers to borrowings: words of the Swadesh lists are seldom borrowed.

The situation seems more complicated if the very fact of genetic affinity has not been proven. Glottochronology as such has not been designed to prove relationships between languages. However, if a sufficient number of similar words is found in the diagnostic lists obtained for the two languages, a genetic relationship seems probable. On the other hand, if two general lexicons share many vocabulary items while the respective diagnostic lists present no or few cognates (as, for instance, in Chinese and Japanese), the respective languages would probably be unrelated, and their shared vocabulary would be attributed to borrowing. Such judgements, however, are largely subjective, since it is unclear what "many" means with respect to shared vocabulary.

The availability of comprehensive data on a language family does not necessarily imply that specific methods of linguistic dating would be worthless. Thus, the groups within the Indo-European family are still listed indiscriminately by comparative linguists, as if they had all formed simultaneously. So far, no conventional gradual classification of the groups has been proposed; it is even unclear whether it should be a tree or a wave-type classification. The solutions could be facilitated by precise data on the age of each group specifying the period of its independent development.

Glottochronology starts from the assumption that each language family derives from one parent source (proto-language). It has also been claimed that a language family can develop from a Sprachbund or result in some other way of convergence of originally unrelated languages. Were it so, glottochronology would be obviously pointless, for it is absurd to date events which have never occurred (the split of a non-existent language family). However, glottochronological reckoning has never yielded any results in favor of the convergence theory.

The Swadesh method seems to have noticeably more opponents than proponents.

Many of the criticisms of, and objections, to glottochronology are naive and do not seem too serious. To cite or discuss all of them would be impossible; here, I will give just one recent statement illustrating this type of criticism. "Languages and their vocabularies develop unevenly, which constitutes their originality and charm"<sup>1</sup>. Some references to specific language data can also be subsumed under this category of judgement. Thus, according to a number of statements, the glottochronological technique, applied to Japanese and Korean, would point to their genetic relationship, for the respective vocabularies share many common items borrowed in parallel from Chinese<sup>2</sup>. In fact, the cognates obtained on the standard diagnostic list will be very few, probably less than five. The review of <sup>2</sup> states, comparing Swadesh lists for Armenian and English would show that these languages are not genetically related<sup>3</sup>; meanwhile, according to G.B. Dzaukjan (1969), the two languages have 20 cognates on the 100-word list, which is close to the English-Russian percentage of cognates.

Further, some objections concern technical problems of glottochronology; for instance it has been asked what should be done if one word in language A has two equivalent words in B (cf., for example, Russian *zola* and *pepel* both denoting 'ashes'). Such difficulties (as well as recommendations concerning possible solutions) do exist,

---

<sup>1</sup> Voprosy jazykoznanija 2 (1984): p.17.

<sup>2</sup> Problema obščnosti altajskix jazykov. Leningrad, 1971, p 53-54.

<sup>3</sup> Voprosy jazykoznanija 4 (1973): p.140.

yet, they cannot destroy general confidence in the method.

Glottochronology has been criticized for dealing exclusively with vocabulary and ignoring phonology and grammar. However, the phonological inventory of any language is relatively small for a statistical method; as to the grammar, very few structural characteristics could be identified as more or less universal. Sound and grammatical changes have not been proved to obey the statistical regularities observed in the vocabulary change (in theory, just the opposite could be assumed). Thus, the employment of phonological and structural data could diminish the utility of the method, rather than increase it.

Turning to serious criticism of glottochronology, two kinds of objections can be distinguished, mathematical and linguistic.

The mathematics of glottochronology has been criticized in most detail by C.D. Chretien, whose main objection to Swadesh's formula is that it is applicable to non-discrete values only. Meanwhile, lexical changes are discrete, for the number of replaced words can only be an integer. This situation has to be represented by another model (binomial, rather than normal distribution). C.D. Chretien presents some calculations done by the alternative technique. For 25 or more cognates (resp. 4500 years of divergence), the "Chretien measurement" appears to be higher than the "Swadesh measurement", with the difference up to three or four centuries; in case less than 25 cognates are observed, the former measurement proves radically lower. It should be emphasized that linguistic dates are inevitably approximate, with the error measured by centuries (see above). The calculation proposed by Chretien is very sophisticated, though its results are generally similar to those obtained by means of the simpler formula

Another objection made by C.D. Chretien (and, earlier, by G. Gleason) deals with the index of retention. It has been shown above that  $r$  is an average for all the items on the diagnostic list. Actually, the list is not homogeneous, for individual elements that form it have different persistence. Less persistent items are substituted on the average at a relatively higher rate; consequently, the index of retention valid for the rest of the list will gradually grow. In the light of this, the  $r$  value, used in



glottochronological measurements, should be considered as a variable (rather than a constant) which slowly changes depending on  $t$ .

If the heterogeneity of the diagnostic list is overlooked, the calculated divergence time will always be lower than the actual value. Distinguishing several groups of elements, differing in persistence, and even calculating the retention rate for every individual item on the list could be a possible solution (actually, this solution has been suggested before). However, it would also lead to sophisticated calculations. For practicing glottochronologists it seems sufficient to be aware of the fact that glottochronological dating will always be somewhat lower than the actual divergence time.

If the mathematics employed in glottochronology is erroneous, while the basic postulates are true (i.e. if the rate of vocabulary change is more or less constant for all languages), the  $t$  value, as obtained by Swadesh's formula, should be regarded as a relative, rather absolute method of dating. Accordingly, glottochronological measurements can be understood as indicating earlier/later divergence time. In this case the method still provides valuable information on the history of language families, though  $t$  values cannot be correlated directly to extralinguistic (historical or archeological) data.

The situation shown is not absolutely hopeless. Making additional, though quite sophisticated, calculations or correlating conditional  $t$  values with actual historical dates, the linguist can draw up correspondences between the relative and absolute time. Though this seems quite a challenge, it is a possible solution.

It was shown above that actual errors in dating due to incorrect formulas, are not very serious (in fact, they are in the same order as the inevitable statistical error).

The linguistic argument against glottochronology focuses upon the statement concerning a constant rate of vocabulary change.

It has been shown by K. Bergsland and H. Vogt (1962) that some languages, documented for a thousand and more years, have retained the basic vocabulary practically intact (e.g. Icelandic). The assumption of constant lexical change seems to be invalid for languages with a consciously developed literary standard. In the

functioning of such languages, mass literacy makes accidental and chaotic substitutions almost impossible. Originally this factor went unnoticed, for the index of retention was determined on the basis of the cases where: a) the literacy tradition was discontinued (as in the comparison of French and Latin), b) the literary standard and the vernacular were separated, developing independently (as in Chinese).

The index of retention was determined on the basis of modern languages. Some theories have been offered that ancient languages were characterized by a different rate of change, being either more conservative or, on the contrary, very flexible and unstable. This opinion is refuted by a series of data; glottochronological measurements obtained for language families prove close to the conventional dates. Thus, modern Indo-European languages demonstrate up to 40 words on the standard diagnostic list dating back to Proto-European (words derived from Indo-European roots with a different original meaning are not included here). Accordingly, Proto-Indo-European can be referred to the epoch of 6000 years ago, which generally accords with the conventional estimate.

Another linguistic objection has been that the separation of dialects is not necessarily followed by independent development. If the related languages neighbor with each other, and their contact and influence are favored by the general environment, the actual lexical differentiation will be lower than it might be expected. Such is the situation with Slavic, and also many Romance and Turkic languages.

Thus, the main statements of glottochronology are not valid for some cases. However, since the situations where glottochronology is unacceptable are known to linguists, this would only mean that the limitations of the method have to be borne in mind

It is worthy of mention that both mathematical and linguistic inadequacies of glottochronology affect the accuracy of measurements in a predictable way: the calculations point to a shorter time (as compared to the actual time). The error is partly canceled by C.D. Chretien's correction, for his formula yields more realistic results for the periods up to 4500 years. The time calculated by Swadesh's formula can, therefore, be interpreted as the minimum possible time of divergence, rather than

the actual or probable time. Incidentally, this repeats M. Swadesh's own estimate of his formula. This approach, however, seems over-cautious to me.

Glottochronology yields the most realistic results for the divergences period of 1500-5000 years. Earlier dates are of little value, for the error can reach 1000 or more years. The method can scarcely be applied to languages with only 5-10 cognates (the respective time of divergence ranges between 8000 and 10,000 years). However, hardly any language families have been attested which split earlier than 6000-8000 years ago. Conventional comparative methods also encounter much difficulty when applied to languages with few cognates; sometimes they simply fail to establish relationships between them.

It would be desirable to apply glottochronology among all established and tentative language families. As a result, language groups could be revealed with a maximum divergence of 60-80 (or, probably, 80-100) centuries, as well as language isolates beyond such groups. Also, realistic and comparable classifications could be proposed for each group.

Translated by Maria Polinsky



## METHODOLOGY OF LONG-RANGE COMPARISON<sup>1</sup>

Sergei Starostin

### 1. Is there any need for long-range comparison?

The question is not uncommon. Some people (even prominent specialists in particular language families) think that they can learn nothing from the outside world and are quite content with what is available.

However, there are two main reasons which, in my opinion, justify the existence of this branch of linguistics:

a) We need to have some classification of the world's languages. The traditional classification (which lists several hundred linguistic families) is a perpetual challenge for comparative linguists. Are there any genetic links between at least some of the world's major linguistic families? If not, how did this extremely strange situation arise? As far as I know, nothing of the kind exists in other disciplines dealing with *Homo Sapiens*, e.g., in the biological sciences.

b) Comparative linguistics is at this time one of the very few branches of science which can supply information about the preliterate history of man. There have been several attempts to combine linguistic data with archeological and genetic evidence, some of which have given very promising results. Surely, if we could extend linguistic evidence to dates earlier than the 4<sup>th</sup> - 5<sup>th</sup> millennia B.C., this could be very useful for the whole field of human history.

### 2. Comparison and reconstruction

The method used by the best long-range comparative linguists was not invented especially for this kind of research. It is the same traditional comparative method

---

<sup>1</sup> This paper has been published earlier in V. Ševoroškin (ed.) 1992: Nostratic, Dene-Caucasian, Austric and Amerind, Bochum, Brockmeyer, p. 75-79.

which has been used in linguistics for nearly two centuries.

There exists, however, a difference - not a methodological, but rather a strategic one: traditional comparative linguistics relies basically on written and spoken languages, whereas the basic material for long-range comparisons is reconstructions.

Of course, the idea of reconstruction is a legitimate part of the traditional comparative method. However, in very many cases, when languages are closely related, genetic classification and different kinds of comparative research are quite possible without any reconstruction. One does not really need a reconstruction to arrive at the idea that, e.g., Slavic languages are related to each other genetically. No reconstruction was needed in the initial stages, when the idea of the Indo-European family was born. In these cases, reconstruction may either be absent altogether (there still exists a large number of commonly accepted linguistic families with no available proto-language reconstruction), or it may be there just as a means of explaining the similarities and correspondences between languages.

For long-range comparison, reconstruction is absolutely vital. One often hears from critically-minded people that if two languages exist separately for a time span of more than 5-6 thousand years, they may lose all traces of similarity and any comparison becomes impossible. They forget, however, that one may deal not with modern languages, but with reconstructed intermediate stages which - for all practical and theoretical reasons - must have been closer to each other than their modern descendants. A few examples:

Modern Chinese numerals *ér* 'two', *wū* 'five' and *bā* 'eight' are totally unlike the Modern Burmese numerals *ne*, *ṇa* and *ṣi*?. However, if we compare reconstructed Old Chinese *\*nij-s* 'two', *\*ṇaʔ* 'five' and *\*prǝ* 'eight' with reconstructed Tibeto-Burman *\*g-nis* 'two', *\*ṇaʔ* 'five' and *\*p-riat* 'eight', we get a fairly good idea of the languages' relationship.

Russian *слышать* 'to hear' (or Old Indian *ṣru-* id., or English *loud*) are certainly not similar to Kor. *kwi* 'ear' or Turkish *kulak* id., or Evenki *ū l-ta-* 'to be heard, resound'. But the reconstructed Proto-Indo-European *\*k'leu-* 'to hear' is much

closer phonetically to reconstructed Proto-Altaic *\*k'üjla* 'ear, to hear'.

Chechen *dog* 'heart' (or Agul *jirk* id., Circassian *g"ə* 'heart, breast') do not resemble the Chinese *yi* 'breast' (or Burmese *raŋ* id.). If we know, however, that the Caucasian forms go back to Proto-North-Caucasian *\*jerkwi*, and the Sino-Tibetan forms - to Proto-Sino-Tibetan *\*ʔək* / *\*ʔəŋ*, the comparison becomes much more plausible.

### 3. Statistical methods

Statistics is not widely used in traditional comparative linguistics. However, it is an important tool for long-range comparison for several reasons:

a) Statistical methods are good for verifying hypotheses about linguistic relationship. Since in many cases long-range genetic links are not superficially obvious, statistical testing is useful for distinguishing between genuine relationships and look-alikes of massive borrowings.

b) Subgrouping in comparative linguistics is usually done using the criterion of shared innovations. In practice, this criterion works best on morphological data. Since the morphological reconstruction of macrofamilies is basically in an initial stage, there is an urgent need for a substitute.

It can be shown that the lexicostatistical method of classifying languages can be applied both to "short-range" and long-range comparison. Since the results obtained in the classification of closely related languages generally correlate rather well with traditional subgroupings, one can assume that the results of long-range classification are also plausible.

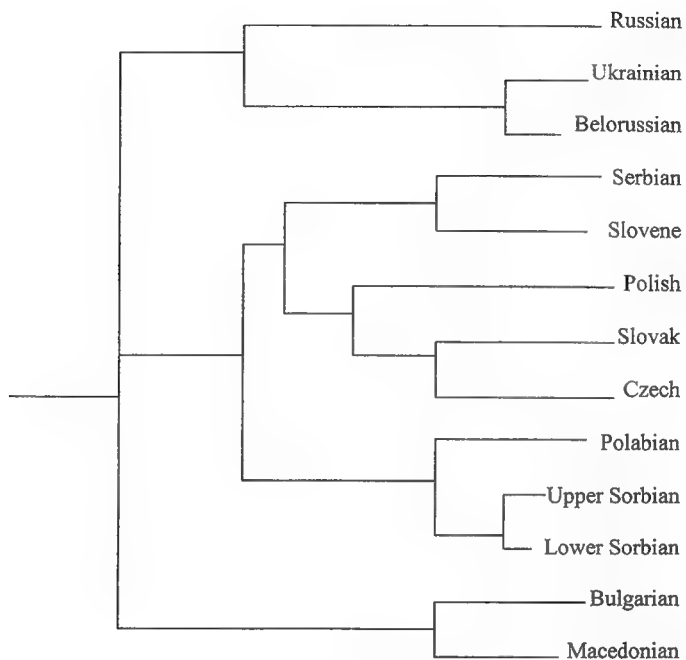
c) The application of statistic methods to linguistic dating (also known as glottochronology) has been widely criticized. While doing "short-range" comparison, one can generally dismiss it and guess the approximate dates of divergence using other evidence (oldest written records, sometimes archeological data). It is, however, the only method which can be applied to distant relationships and, therefore, seems to be worth re-examining.

#### 4. Computer methods

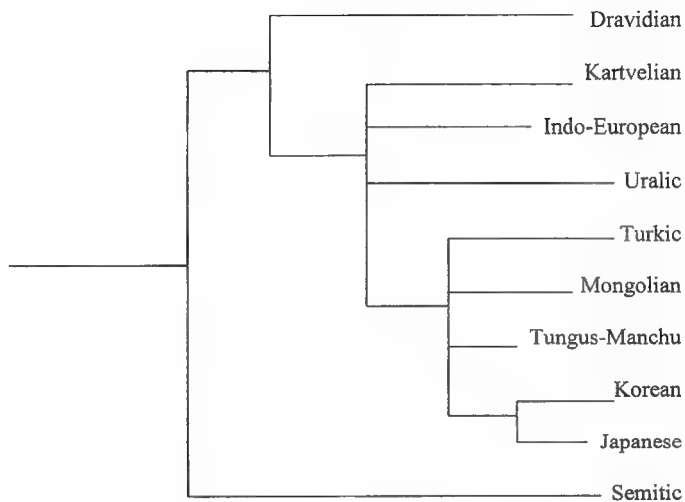
A researcher dealing with long-range comparison has to process a huge amount of linguistic data, which grows exponentially once any new linguistic family is being added. Modern computer technology allows one to deal with this flow of data more efficiently, although there still are very few computer applications designed for comparative linguistics. It is possible to use computers for storing large comparative databases, for processing data of related languages (even for establishing phonetic correspondences), and, of course, for performing all kinds of statistical calculations.



A Computer Generated Genealogical Tree for Slavic



A Computer Generated Genealogical Tree for Nostratic



## SOME NOTES ON LINGUISTIC COMPARISON

Alexander Vovin

The goal of this article is to discuss certain methodological aspects of language comparison in general and long-range comparison in particular. I intend to demonstrate that the traditional comparative method is self-sufficient for establishing genetic relationships for languages belonging to families consisting of languages more remotely related than, let us say, Indo-European, or between language families themselves. First, I will address the issue of the so-called "Omnicomparativismus" (Doerfer 1973), often raised by the opponents of long range comparison. There is widespread opinion that, relying exclusively on the comparative method, it is possible to take any two languages and demonstrate that they might be remotely related. I believe that this point of view has actually a religious rather than scholarly basis: to the best of my knowledge there has never been a successful attempt at applying the comparative method to two or more unrelated languages and ultimately proving that these languages are remotely related. Moreover, it seems to me that in reality nobody has ever tried to prove that it is indeed possible to take any two languages and, using the comparative method, demonstrate that they might be remotely related. Therefore, this point of view has doubtful scholarly value and represents an ad hoc approach. I will demonstrate exactly the opposite: using the comparative method it is possible to prove that two arbitrarily chosen languages are unrelated. Let us try to prove that Japanese (Japanese-Ryukyuan group, Japanese islands) and Hopi (Uto-Aztecan family, Arizona) are related. For this purpose I will compare M. Swadesh 100-word lists for Japanese and Hopi<sup>1</sup>, will try to identify some

---

<sup>1</sup>I wish to express my gratitude to Alexis Manaster-Ramer, Wayne State University, for his help with Hopi and Proto-Uto-Aztecan data I used (Miller 1967, 1987), (Albert & Shaul 1985), (Kalectaca 1982) for lexical data.

look-alikes within that list, and then demonstrate that these look-alikes are, indeed, cognates, that is, they are based on regular phonetic correspondences existing between Japanese and Hopi.

<u>gloss</u>	<u>Japanese</u>	<u>Hopi</u>
all	minna, subete	soosoy(am), sokyawat
ashes	hai	qōtsvi
bark (n.)	kawa	qaapu
belly	hara	pono
big	ooki-	wuuko
bird	tori	tsiro
bite	kam-	kuuki
black	kuro-	qōm(a)vi
blood	ti	ungwa
bone	hone	ōōqa
breast	tibusa, mune	pīi-hu
burn	yak-	tawtsikya
cloud	kumo	ooma-wu
cold	samu-	suusungwa, iyoho'o
come	ko-	pew'i
die	sin-	aapiy, mooki
dog	inu	pooko
drink	nom-	hiiko
dry	kawak-	lakna, mōōqa
ear	mimi	naqvu
earth	tuti	tutskwa, tuuwaqatsi
egg	tamago	nō-hu
eat	tabe-, kuw-	nōōsa, tuumoyta
eye	me	poosi
fat	abura, siboo	wihutoyna

# SOME NOTES ON LINGUISTIC COMPARISON

feather	hane	homasa, masa
fire	hi	qōōhi, qōhō, uuwingwa
fish	uo, sakana	paakiw
fly	tob-	puuyalti
foot	asi	kuku
full	ippai, miti-	oopo-
give	atae-	maqa
go	ik-	-ma
good	yo-, i-	nukngwa, lolma, nukwangw-
green	ao-, midori	sakawawsa, mokingpu
hair	ke, kami	pōhō (body), hōōmi (head)
hand	te	ma-
head	atama	qōtō
hear	kik-	tuuqayi, navota
heart	sinzoo, kokoro	unangwa
horn	tuno	aala
I	watakusi	nu'
kill	koros-	niina (sing.), qōya (plur.)
knee	hiza	tamō
know	sir-	navota, tuwi'[yta]
leaf	happa, ha	naapi
lie	yoko ni nar-	wa'ō
liver	kanzoo, kimo	nuuma
long	naga-	wuupa
louse	sirami	atu
man	otoko, dansei	taaqa
many	oo-	soo-, kyasta, naavinta, wuuhaq, kyaysiwa
meat	niku	sikwi, nōq-

moon	tuki	muuyaw(u)
mountain	yama	tuukwi
mouth	kuti	mo'a
nail	tume	mu'aapi
name	namae	tungwni
neck	kubi	kwaapi
new	atarasi-	puuhu
night	yoru	tookila
nose	hana	yaqa
not	-ana-, na-	qa, so'on
one	iti, hito-	suukya
person	hito	sino
rain	ame	yooyangw
red	aka-	pala-
root	ne	ngahu
round	maru-	pongokpu, pōlangpu
sand	sunā	tuuwa
say	iw-	pangawu, hingqawa, kita, lavayti
see	mi-	tuwa
seed	tane	poosi, poshumi, sivosi
short	mizika-	tšaava
sit	suwar-	poni, qatuptu, atsviewta
sleep	ne-	puwta
small	tiisa-	hisay-, tšaako, tšaayo, -hoya
smoke	kemuri	kwiitsingwu
stand	tat-	wunuwta
star	hosi	soohu
stone	isi	owa
sun	tayoo, hi	taawa

swim	oyog-	momori
tail	sippo, o	suru
that	so-, a-	mi'
this	ko-	i'
tongue	sita	lengi
tooth	ha	tama
tree	ki	-tsoki
two	ni, huta-	lööy[ö]-, lööq
warm	atataka-	yongi
water	mizu	paa-hu
way	miti	pö-hu
we	ware-ware, watakusi-tati	itam
what	nani	himu
white	siro-	qöötsa
who	dare	hak
woman	onna, zyosei	wuuti, tumasi
yellow	kiiro-	sikya
you	kimi, omae, anata	um, uma

Comparing Japanese and Hopi lists, we can find the following eight look-alikes: *J kawa*, *H qaapu* "bark"; *J ooki-*, *H wuuko* "big"; *J tori*, *H tsiro* "bird"; *J kami*, *H höömi* "hair"; *J tuti*, *H tutskwa* "earth"; *J kubi*, *H kwaapi* "neck": *J niku*, *H nöq-* "meat"; *J taiyoo*, *H taawa* "sun". Unfortunately for the purpose of proving that Japanese is related to Hopi, the two last words in Japanese are loanwords from Chinese: *J niku* < Middle Chinese (MC) *ñiwk* "meat", *J taiyoo* < MC *taiyang* "sun". If we further use Old Japanese (OJ) forms: *kapa* "bark", *opo-* "big", *tori* "bird", *kami* "hair", *tuti* "land", then we also have to exclude the comparison of OJ *opo-* and *H wauto* "big" as unlikely. Therefore, we are left with four "cognates". Let us try to establish a system of regular phonetic correspondences on their basis.

<u>OJ</u>	<u>H</u>
k	q
k	kw
a	aa
u	aa
p	p
b	p
t	ts
<u>q</u>	i
i	o
k	h
a	öö
m	m
yi	i
l	t

The system above can hardly be called a regular one: thus, OJ /k/ corresponds to H /q/, /kw/, and /h/, OJ /t / corresponds to both H /t/ and /ts/, OJ /a/ corresponds to both H /aa/ and /öö/, H /i/ corresponds to both OJ /q/ and /yi/. An attempt to find other look-alikes in Hopi which would support the same correspondences also produced a negative result. Therefore, a low percent of look-alikes within the 100 basic vocabulary list, and lack of regular correspondences between these look-alikes suggest that we are dealing with two unrelated languages. However, before we jump to this conclusion, let us compare same basic word list, compiled for Proto-Japanese and Proto-Uto-Aztec. If our assumption that Japanese and Hopi are unrelated is wrong, then we should expect a certain increase in the number of possible cognates. Moreover, we might be able to establish regular phonetic correspondences between these possible cognates.



<u>gloss</u>	<u>PJ<sup>2</sup></u>	<u>PUA</u>
all	*múCí-nà	?
ashes	*pápí	*nasi
bark (n.)	*kàpà	*ko
belly	*pàrà	*sapu
big	*òpò-	*we, *k <sup>w</sup> e, *pi(y)a
bird	*tóri	*wici, *wiki, *cutu
bite	*kàm-	*ke, *keyV
black	*kùrwò	*tu, *tuhu, *cuk
blood	*ti	*'etV, *'ewV
bone	*pone	*'o, *oho
breast	*ti/*titi	*pi, *tawi
burn	*dák-	*na, *nai
cloud	*kùmù[C]à	*to, *top, *tom
cold	*sàmù-	*se, *sep
come	*kò-	*kim
die	*sin-	*muk[i] (sing.), *koi (pl.), *su[w]a (pl.), *te
dog	*inù	*puku, *cu
drink	*nòm-	*hi, *hi'e
dry	*káv(V)rà-k-	*waki
ear	*mimi	*na(N)ka
earth	*tùti	*te-
egg	*kwo	*no
eat	*kup-	*k <sup>w</sup> a
eye	*mà-n	*pusi
fat	*à(n)pùrá	*wi
feather	*pánĚ	*pi

---

<sup>2</sup>In order to save space PJ high pitch is rendered by acute and low pitch by gravis.

fire	*pò-Ci	*ku
fish	*(d)iwó	*-ŋkʷi, *top
fly	*tónp-	*ya, *hini, *ne
foot	*pàŋki	*ta, *to, *nape, *ke, *keke
full	*mìt-	*pu
give	*ata[-]pa-Ci	*maka
go	*kádwóp-, *dik-	*miya, *simi
good	*dò-	*cam, * 'ay
green	*àwò, *míntóri	---
hair	*ká-Ci	*suwi, *pe (body), *kupa (head) *wo (head)
hand	*tà-Ci	*ma
head	*tumu-, kàsirà	*mo'o, *co, *kat
hear	*kí[-]k-	*ka
heart	*kòkòró	*pi(h)wi, *pi(h)yi, *sula
horn	*tùnwò	* 'awa
I	*bàn[u]	*ni
kill	*kórós-	*paka, *paki
knee	*pínsá, PR *tubusin	*tona
know	*sír-	*mati, *maci
leaf	*pá	*sawa
lie	*ná-	*po
liver	*kímwò	*nema
long	*nànkà-	*tepV, *te
louse	*sirámì	* 'ate
man	*bò	*taka
many	*mana-Ci	*yo, *mui, *mu'i
meat	*sisi	*wa'i, *tuhku
moon	*tùkú-	*meya

# SOME NOTES ON LINGUISTIC COMPARISON

mountain	*dàmà	*kawi
mouth	*kútú-Ci	*teni
nail	*túmá-Ci	*sutV
name	*ná	*tewV
neck	*kúnpi	*kuta
new	*àrà-ta-	*pe
night	*dùCà	*tukV
nose	*páná	*yaka
not	*-an[a]-	*ka, *kai
one	*pitò	*se
person	*pítò	*tewi
rain	*àmâ-Ci	*yukV (v.), *(w)ema (v.)
red	*áká-	*set
root	*mòtò, *nE	*na
round	*márú/*máró	*pot
sand	*súná	*si, *se, *se
say	*ip-	*ya
see	*mì-	*te[w?]
seed	*táná-Ci	*pusi, *paci
short	*m-insikà-	*tup
sit	*bí-	*ya, *yas, *kate (sing.)
sleep	*ui-	*pei, *ku, *kup
small	*tipisà-	*'ali, *te, *no
smoke	*kái[-]npúri	*k <sup>w</sup> i, *kuhi
stand	*tát-	*wene, ?*wele, *ke
star	*pósí	*su, *cu
stone	*(d)ísò	*te
sun	*pí	*ta, *tape
swim	*òyò-	---

tail	*b <sub>Q</sub>	*k <sup>w</sup> asi, *k <sup>w</sup> aci
that	*ká-, *a-	*p <sub>θ</sub>
this	*k <sub>Q</sub> -	*'i
tongue	*sità	*neni
tooth	*pà	*taSma, *tamaN
tree	*k <sub>Q</sub> -< *k <sub>Q</sub> or	*ku
two	*puta	*wehV "two"
warm	*àta-taka-	*yu
water	*mí	*pa
way	*míti	*po
we	*bàn[u]	*ta(h)-mV, *ta(h)-nV
what	*nà[-]ní	*hii
white	*sírà-Cu	*tosa
who	*tá-	*haki, *hake
woman	*-mina/*míCá	*su/*so(n), *hupi, *nawi
yellow	*kú-Ci	*si, *ci, *'oha/*'uha, *sawa
you	*si/*s <sub>Q</sub> -, *na	*ih, *iN

We can again probably find five suspects: PJ \*kàpà, PUA \*ko "bark"; PJ \*k<sub>Q</sub>-, PUA \*kim "come" (cf. English come?!); PJ \*kup-, PUA \*k<sup>w</sup>a "eat"; PJ \*ki[-]k-, PUA \*ka "hear"; PJ \*duCa, PUA \*tuk "night". It is worth noting that with the exception of "bark" all these words are different from the look-alikes in the Japanese-Hopi list. Besides, the actual number of possible "cognates" actually decreased almost twofold: from eight to five. A brief list of correspondences will easily demonstrate that there is no regularity in them and that every suggested comparison is based on similarity in one phoneme:

# PJ      PUA

*k	*k
*a	*o
* <sub>Q</sub>	*i

*-Ø	*-m
*k	*k <sup>w</sup>
*u	*a
*p	*Ø
*i	*a
*d	*t
*u	*u
*C	*k

PJ \*k corresponds to both PUA \*k and \*k<sup>w</sup>; PUA \*k corresponds to both PJ \*k and \*C; PJ \*u corresponds to both PUA \*u and \*a; PUA \*a corresponds to both PJ \*u and \*i. Therefore, our assumption that Japanese and Hopi are unrelated languages is supported further by the comparison of PJ and PUA, and we can arrive at the definite conclusion that these languages are not related.

Now, let us evaluate the genetic relationship between Japanese and Turkish in the same way we did with Japanese and Hopi. Turkish is located in Asia Minor and belongs to the Turkic group of languages. It gives us, if not the same distance between Japanese and Turkish as between Japanese and Hopi, but at least somewhat comparable. Like the case with Japanese and Hopi, there are no historically attested contacts between the Japanese and Turkic languages. However, while some linguists, among them the author of these lines believe that Japanese and Turkic languages are related and that both are members of the Altaic language family, comprising besides Japanese and Turkic also Mongolian, Manchu-Tungus, and Korean (Miller 1967, 1971, 1980), (Menges 1975), (Murayama 1957, 1962), (Starostin 1986, 1991), (Street & Miller 1975-1978), (Vovin 1993a, 1993b, 1994); there are also opponents of the theory (Doerfer 1974), (Janhunen 1992), (Comrie 1993). In order to resolve this controversy, let us apply the test to Turkish and Japanese. If they fail it, like Japanese and Hopi did, then Japanese and Turkish are not related. However, if they pass it, then these two languages must be related. Therefore, let us compare 100-word lists for Japanese and Turkish:

<u>gloss</u>	<u>Japanese</u>	<u>Turkish</u>
all	<b>minna</b> , subete	<b>bütün</b> , hep
ashes	hai	kül
bark(n.)	<b>kawa</b>	<b>kabuk</b>
belly	hara	karın
big	ooki-	büyük
bird	tori	kuş
bite	kam-	dişle-, tıstır
black	<b>kuro-</b>	<b>kara</b> , siyah
blood	ti	kan
bone	hone	kemik
breast	tibusa, mune	göğüs
burn	<b>yak-</b>	<b>yak-</b>
cloud	kumo	bulut
cold	samu-	soğuk
come	<b>ko-</b>	<b>gel-</b>
die	sin-	öl-
dog	inu	köpek, it
drink	nom-	iç-
dry	<b>kawak-</b>	<b>kuru-</b>
ear	mimi	kulak
earth	tuti	yer
egg	tamago	yumurta
eat	tabe-, kuw-	ye-
eye	me	göz
fat	abura, siboo	yağ
feather	hane	tüy
fire	hi	ateş
fish	uo, sakana	balık

fly	tob-	uç-
foot	asi	ayak
full	ippai, miti-	dolu
give	atae-	ver-
go	ik-	git-
good	yo-, i-	iyi, <b>yahşi</b>
green	ao-, midori	yeşil
hair	<b>ke</b> , kami	<b>kıl</b> , saç
hand	te	el
head	atama	baş
hear	kik-	duy-, işit-
heart	sinzoo, kokoro	yürek
horn	tuno	boynuz
<b>I</b>	<b>watakusi</b>	<b>ben</b>
kill	koros-	öldür-
knee	hiza	diz
know	sir-	bil-
leaf	happa, ha	yaprak
lie	yoko ni nar-	yat-
liver	kanzoo, kimo	karaciğer
long	naga-	uzun
louse	sirami	bit
man	otoko, dansei	erkek, adam
many	oo-	çok
meat	niku	et
moon	tuki	ay
mountain	yama	dağ
mouth	kuti	ağız
nail	<b>tume</b>	<b>tırnak</b>

name	namae	at (ad-), isim
neck	kubi	boyun
new	atarasi-	yeni
night	yoru	gece
nose	hana	burun
not	-ana-, na-	-ma-, değil, yok
one	iti, <b>hito-</b>	<b>bir</b>
person	hito	insan
rain	ame	yağmur
red	aka-	kızıl, kırmızı,
root	ne	kök
round	maru-	yuvarlak, değirmi
sand	suna	kum
say	iw-	de-
see	mi-	gör-
seed	tane	tane, tohum
short	mizika-	kısa
sit	suwar-	otur-
sleep	ne-	uyu-
small	tiisa-	küçük
smoke	kemuri	duman
stand	<b>tat-</b>	<b>dur-</b>
star	<b>hosi</b>	<b>yıldız</b>
stone	<b>isi</b>	<b>taş</b> < Old Osman <b>daş</b>
sun	taiyoo, hi	güneş
swim	oyog-	yüz-
tail	sippo, o	kuyruk
that	so-, a-	şu, o
this	ko-	bu



# SOME NOTES ON LINGUISTIC COMPARISON

tongue	sita	dil
tooth	ha	diş
tree	ki	ağaç
two	ni, huta-	iki
warm	atataka-	sıcak
water	mizu	su (suy-)
way	miti	yol
we	<b>ware-ware, watakusi-tati</b>	<b>biz</b>
what	<b>nani</b>	<b>ne</b>
white	siro-	ak, beyaz
who	dare	kim
woman	onna, zyosei	kadın
yellow	kiiro-	sarı
you	kimi, omae, anata	sen (sing.), siz (plur.)

In the chart above I gave in bold typeface the words which we Altaicists consider to have the same origin in Japanese and Turkish. A reader unfamiliar with histories of both Japanese and Turkic may be surprised that there are such dissimilar words on the list as J *isi* "stone" and T *taş* "id", or J *minna* "all" and T *bütün* "id", while such look-alike words as J *tane* "seed" and T *tane* "id" are not included. T *tane* "seed" violates the vowel harmony and therefore induces a suspicion that it is a loanword, which it indeed is: from Persian. The cognates, on the other hand, are established on the basis of regular phonetic correspondences, which I will provide further below.

The system of consonantal correspondences<sup>3</sup> between Japanese and Turkish will be following:

---

<sup>3</sup>Vowel correspondences between Japanese and Turkish are considerably more complicated than consonantal correspondences. This situation is by no means unique for the Altaic family: the lack of one-to-one vowel correspondences is observed in all pretty well established families: Uralic, Austroasiatic, Afroasiatic, and even Indo-European.

<u>Japanese</u>	<u>Turkish</u>
h-, -w-/-Ø-	Ø-, b
w-/ ___a, Ø	b
m	b-, m
t	t
t/y, Ø/ ___i	d
y, Ø/ ___i	y
n	y
-t-, -r-, -Ø	-r-
-r-, -Ø	-l-
-s-	-š-
t	ç
s	s
k	k/g
-m-	-n-
-n-	-t-

Some of these correspondences may seem too complicated or even unnatural, but they will be fully explained further.

Japanese and Turkic occupy a unique position within Altaic in the sense that both are attested by a considerable number of texts earlier than those in other Altaic languages: from the beginning of the eighth century. Therefore, it might be interesting to have a look at the corresponding list for Old Japanese and Old Turkic: if the languages in question are related we should expect a slight increase in the number of cognates within the 100-word list. This is indeed what happens: we have 21 cognates in the following list versus 17 in the first:

<u>gloss</u>	<u>Old Japanese</u>	<u>Old Turkic</u>
all	miyna	bütün, qop
ashes	papi	kül
bark (n.)	kapa	qabuq

# SOME NOTES ON LINGUISTIC COMPARISON

belly	para	qarīn
big	opo-	beʔük
bird	tōri	quš
bite	kam-	īsīr
black	<b>kurwo-</b>	<b>qara</b>
blood	ti	qan
bone	pone	sūnjūk
breast	mune (muna-)	kōküz
burn	<b>yak-</b>	<b>yaq-, örte-</b>
cloud	kumwo	bulut
cold	samu-	soyuq
come	<b>kq-</b>	<b>kel-</b>
die	sin-	öl-
dog	<b>inu</b>	<b>īt</b>
drink	nōm-	ic-
dry	<b>kawak-</b>	<b>quruq,</b>
ear	myimyi	qulqaq
earth	tuti	yer
egg	kwo	yumurtya
eat	kup-	ye-
eye	mey	köz
fat	abura	yay
feather	pane	yüg
fire	<b>piy (pō-)</b>	<b>ot, ört</b>
fish	iwo	balīq
fly	tōb-	uc
foot	asi	aʔak
full	mit-	tolu
give	atae-	ber-

go	ik-	ket-
good	<b>yo-</b>	<b>yaqšī, yäg, ädgü</b>
green	awo-, myidori	yasıl
hair	<b>key (ka-), kamyi</b>	<b>qıl, sac</b>
hand	te (ta-)	elig
head	kasira	baş
hear	kyik-	ešit-
heart	<u>koko</u> ro	yürek
horn	tunwo	müñüz
I	<b>wa-</b>	<b>ben</b>
kill	<u>koro</u> s-	ölür-
knee	pyiza	tiz
know	sir-	bil-
leaf	pa	yapuryaq
lie	ne-, pus-	yat-
liver	kyimwo	bayır
long	naga-	uzun
louse	sirami	bit
man	<u>woto</u> ko	er
many	mane-	köp, üküš
meat	sisi	et
moon	tukiy	ay
mountain	yama	tay
mouth	kuti	ayız
nail	<b>tume</b>	<b>tırnak</b>
name	na	at
neck	kubi	boyun
new	arata-	yañı
night	ywo, yworu	tün

nose	pana	burun
not	-ana-, na-	-ma-
one	<b>pyito-</b>	<b>bir</b>
person	pyito	kişi
rain	amey	yaymur
red	aka-	qızıl
root	<b>ne</b>	<b>kök, yiltüz</b>
round	rnaru-	yumyaq, tegirmi
sand	isagwo, suna	qum
say	ip-	te-
see	myi-	kör-
seed	tane	uruy
short	myizika-	qışya
sit	suwar-	otur-
sleep	ne-	uđi-
small	tipyisa-	kicig
smoke	keymuri	tütün
stand	<b>tat-</b>	<b>tur-</b>
star	<b>posi</b>	<b>yulduz/yultuz</b>
stone	isi	taş
sun	pyi	kün, küneş
swim	oyog-	yüz-
tail	wo	quðruq
that	sə-, ka-, a-	<b>an</b>
this	kə-	bu
tongue	sita	ül
tooth	pa	tiş/tiş
tree	kiy (kə-)	ıyac
two	ni, huta-	iki/eki

warm	atataka-	yǎlıy
water	mizu	sub
way	miti	yol
we	<b>wa-</b>	<b>biz</b>
what	<b>nani</b>	<b>ne</b>
white	sirwo-	ak, ürüg
who	dare	kim
woman	onna, zyosei	evci
yellow	kyi	sarıy
you	<b>si (so-), na</b>	<b>sen (sing.), siz (plur.)</b>

The system of consonantal correspondences between Old Japanese and Old Turkic is the following:

<u>Old Japanese</u>	<u>Old Turkic</u>
p	Ø-, b
w	b
m	b-, m
t	t
t/y, Ø/___i	d
y, Ø/___i	y
n	y
-t-, -r-, -Ø	-r-
-r-, -Ø	-l-
-s-	-š-
t	c
s	s
k	k/g
-m-	-n-
-n-	-t-

By the same token, we can expect that the number of cognates will increase

if we further compare Proto-Japanese with Proto-Turkic<sup>4</sup>, and again we have twenty-eight cognates:

<u>gloss</u>	<u>PJ</u>	<u>PT</u>
all	*múCí-nà	*büt-ün
ashes	*pápí	*kül <sub>1</sub>
bark (n.)	*kàpà	*kaapuk
belly	*pàrà	*kar <sub>1</sub> in
big	*òpò-	*bädü-k
bird	*tórí	*kul <sub>2</sub>
bite	*kàm-	*isir <sub>1</sub> -
black	*kùrwò < *kura-Cu	*kar <sub>1</sub> a
blood	*tí	*kaan
bone	*pone	*kämik
breast	*ti/*titi	*emV, *cici
burn	*dák-	*yak
cloud	*kùmù[C]à	*bul <sub>1</sub> ut/*bul <sub>1</sub> it
cold	*sàmù-	*sogĩ-k
come	*kò-	*gäl <sub>1</sub> -
die	*sin-	*öl-
dog	*inù	*it
drink	*nòm-	*ic-
dry	*káv(V)rá-k-	*kuur <sub>1</sub> i-k
ear	*mimi	*kul-kak
earth	*tùti	*yer <sub>1</sub>
egg	*kwo	*yumurtka
eat	*kup-	*yee-
eye	*mà-n	*gör <sub>2</sub>

---

<sup>4</sup>Proto-Japanese reconstruction mainly follows (Martin 1987), and Proto-Turkic follows (Starostin 1991), but both include certain revisions and corrections by the author.

fat	*à(n)pùrá	*yaag
feather	*pánÉ	*tüg/*tük
fire	*pò-Ci	*ör <sub>1</sub> -t, *ot
fish	*(d)íwó	*baal <sub>1</sub> ik
fly	*tónp-	*uc-
foot	*pànkì	*adak
full	*mit-	*dool <sub>1</sub> ĩ
give	*ata[-]pa-Ci	*beer <sub>1</sub> -
go	*kádwóp-, *dik-	*bar <sub>1</sub> -, *geed-
good	*dò-	*yag-, *äd-gü
green	*àwò, *míntórí	*yaal <sub>2</sub> -ĩl <sub>1</sub>
hair	*ká-Ci	*kĩl <sub>1</sub> , *sac
hand	*tà-Ci	*äl <sub>1</sub>
head	*tumu-, *kàsìrà	*bal <sub>2</sub>
hear	*kí[-]k-	*el <sub>2</sub> it
heart	*kòkòró	*yür <sub>1</sub> äk
horn	*tùnwò	*bũñür <sub>2</sub>
I	*bàn[u]	*bän
kill	*kórós-	*öl <sub>1</sub> -dür <sub>1</sub> -
knee	*pínsá, PR *tubusin	*diir <sub>2</sub>
know	*sír-	*bil <sub>1</sub> -
leaf	*pá	*yapur <sub>1</sub> -gak
lie	*ná-	*yat-
liver	*kímwò	*bagĩr <sub>1</sub>
long	*nànkà-	*ur <sub>2</sub> ĩ-
louse	*sirámí	*bit
man	*bò	*äär <sub>1</sub>
many	*mana-Ci	*cok
meat	*sìsì	*ät



moon	*tùkú-	*aañ
mountain	*dàrà, *tàka-Ci	*daag
mouth	*kútú-Ci	*agĩr <sub>2</sub>
nail	*túmá-Ci	*dĩr,ŋa-k
name	*ná	*aat
neck	*kúnpi	*boyun
new	*àrà-ta-	*yaŋĩ/*yegĩ
night	*dùCà	*geecă
nose	*páná	*bur <sub>1</sub> un/*bur <sub>1</sub> ĩn
not	*-an[a]-	*-mV-
one	*pitò	*bir <sub>1</sub>
person	*pítò	*kil <sub>2</sub> i
rain	*àmâ-Ci	*yagmur <sub>1</sub> “falling water”
red	*àkà-	*kĩr <sub>2</sub> -ĩl <sub>1</sub> , *aakV
root	*mòtò, *nE	*yĩl <sub>1</sub> , *kòk, *düüp
round	*márú/*máró	*yum-, *yub
sand	*súná	*kum
say	*ip-, *tò-	*dee-
see	*mì-	*gõr <sub>1</sub> -
seed	*táná-Ci	*ur <sub>1</sub> ug
short	*m-ĩnsikà-	*kĩs-
sit	*bí-	*ol <sub>1</sub> -ur <sub>1</sub> -
sleep	*uĩ-	*uu-dĩ-
small	*tipisà-	*kicik
smoke	*káĩ[-]npúrí	*tüt-
stand	*tát-	*dur <sub>1</sub> -
star	*pósí	*yul <sub>2</sub> -tur <sub>2</sub>
stone	*(d)ísò	*daal <sub>2</sub>
sun	*pí	*gũn[-ăl <sub>2</sub> ]

swim	* <u>ôyô</u> -	*yür <sub>2</sub> -
tail	*b <u>ô</u>	*kudruk
that	*ká-, *a-	*a-, *o-
this	*k <u>ô</u> -	*ku, *bu
tongue	*sità	*dīl <sub>1</sub> /*dīl <sub>1</sub>
tooth	*pà	*dīl <sub>2</sub> , *siVl <sub>1</sub>
tree	*k <u>ô</u> - < *k <u>on</u> or	*i[ŋ]gac
two	*puta	*eki/*iki
warm	*àta-taka-	*yīl <sub>1</sub> ī-g
water	*mí	*sub, *yag-mur <sub>1</sub> "falling water" = "rain"
way	*mítí	*yool
we	*bàn[u]	*bi-r <sub>2</sub>
what	*nà[-]ní	*nV
white	*sírà-Cu	*aakV, *siVr <sub>1</sub> V
who	*tá-	*kim/*kem
woman	*-mina/*míCá	---
yellow	*kú-Ci	*siVr <sub>1</sub> V
you	*sí/*s <u>ô</u> -, *na	*sän

Finally, we can give the set of regular phonetic correspondences between Proto-Japanese and Proto-Turkic (together with their archetypes in Proto Altaic):

<b>PJ</b>	<b>PT</b>	<b>PA</b>
*p	*Ø, p	*p'
*b	*b	*b
*m	*b-, m	*m
*t	*t	*t'
*t/*d	*d	*t
*d	*y	*d
*n	*y	*n

*-t-, -r-, -Ø	*-r <sub>1</sub> -	*-r <sub>1</sub> -
*-r-, -Ø	*-l <sub>1</sub> -	*-l <sub>1</sub> -
* - s-	*-l <sub>2</sub> -	*-l <sub>2</sub> -
*t	*c	*c
*s	*s	*s
*k	*k/*g	*k', *k, *g
*-m-	*- -	*- -
*-n-	*-t-	*-nt-
*u	*ü	
*u	*o	*o
*a	*a	*a

It is also necessary to keep in mind two important phonotactic rules:

- 1) PA \*C<sub>1</sub>VC<sub>2</sub>[+voice] > PJ \*C<sub>1</sub>V, PT \*C<sub>1</sub>VC<sub>2</sub>;
- 2) PA \*C<sub>1</sub>V<sub>1</sub>C<sub>2</sub>V<sub>2</sub> > PJ \*C<sub>1</sub>V<sub>1</sub>C<sub>2</sub>V<sub>2</sub>, PT \*C<sub>1</sub>V<sub>1</sub>C<sub>2</sub>[V<sub>2</sub>].

The charts of correspondences for the daughter languages: Japanese and Turkish, Old Japanese and Old Turkic were given above. I would like to emphasize that it is important to give the system of correspondences for protolanguages because various phonological developments in daughter languages obscure to a certain extent the original systems of the corresponding protolanguages. Therefore, the charts I provided above are more complicated and would require extensive commentaries.

In sum, it is impossible with the traditional comparative method to prove that two arbitrarily chosen languages are related. On the contrary, the comparative method demonstrates that these languages are not related. It is always possible to find several look-alikes within any two 100 word lists. However, if the languages in question are not related, it is impossible to establish any system of regular phonetic correspondences. Moreover, comparing reconstructions of these languages, one can easily demonstrate that potential "cognates", chosen on the look-alike principle, are not cognates at all, and even the number of look-alikes will be less if one compares reconstructions. On the other hand, if any two given languages are indeed related, we

obtain quite the opposite results. First of all, it is possible to establish a system of regular phonetic correspondences. Second, the number of cognates will increase significantly if older forms of languages in question are compared. The increase of a number of cognates becomes even more significant if we compare the reconstructions of related languages.

LITERATURE

- Albert, Roy & Shaul, David L. 1985. *A Concise Hopi and English Dictionary*.  
John Benjamins Publishing Company.
- Comrie, Bernard 1993. *Review of Starostin 1991*.  
In Language 69/4: 828-832.
- Doerfer, Gerhard 1973. *Lautgesetz und Zufall. Betrachtungen zur Omnicomparativismus*.  
In Innsbrucker Beiträge zur Sprachwissenschaft 10. Innsbruck.
- Doerfer, Gerhard 1974. *Ist das Japanische mit den altaischen Sprachen verwandt?*  
In Zeitschrift der Deutschen Morgenländischen Gesellschaft 124:103-142.
- Janhunen, Juha 1992. *Das Japanische in vergleichender Sicht*.  
In Journal de la Société Finno-Ougrienne, 84: 145-161.
- Kalectaca, Milo 1982. *Lessons in Hopi*.  
University of Arizona Press. Tucson.
- Martin, Samuel E. 1987. *Japanese Language Through Time*.  
New Haven & London: Yale University Press.
- Menges, Karl. 1975. *Altaiisch und Japanisch*. (= Abhandlungen für die Kunde des Morgenlandes, 41:3.)  
Wiesbaden: Deutsche Morgenländische Gesellschaft.
- Miller, Roy A. 1967. *The Japanese Language*.  
Chicago & London: Chicago University Press.
- Miller, Roy A. 1971. *Japanese and the Other Altaic Languages*. Ibid.
- Miller, Roy A. 1980. *Origins of the Japanese Language*.  
Seattle & London: University of Washington Press.
- Miller, Wick R. 1967. *Uto-Aztecan Cognate Sets*.  
University of California Publications in Linguistics, 48.
- Miller, Wick R. 1987. *Computerized Data Base for Uto-Aztecan Cognate Sets*.  
Computer Printout. Salt Lake City.
- Murayama, Shichirō. 1957. *Vergleichende Betrachtung der Kasus-Suffixe im Altjapanischen*.  
In Studia Altaica: Festschrift für Nikolaus Poppe zum 60 Geburtstag (= Ural-Altaische Bibliothek, 5), 126-131.  
Wiesbaden: Otto Harrassowitz.
- Murayama, Shichirō. 1962. *Nihongo no tungusugo teki yōso*.  
In Minzokagaku kenkyū 26/3.

- Starostin, Sergej A. 1986. *Problema genetičeskoj obsnosti altaiskix jazykov*.  
In Istorikokul'turnye kontakty narodov altaiskoi jazykovoi obsnosti, vol.II, 105-112.  
Moscow: Izd. Nauka. [Materials from the 29th PIAC Conference.]
- Starostin, Sergej A. 1991. *Altaiskaia problema i proisxoždenie japonskogo jazyka*.  
Moscow: Izd. Nauka.
- Street, John & Roy Andrew Miller. 1975-78. *Altaic Elements in Old Japanese*. 2 vols.  
Madison, Wis.: Authors. [Vol.II by Street only.]
- Vovin, Alexander 1993a. *On the Phonetic Value of the Middle Korean Grapheme Δ 'Triangle'*.  
In BSQAS 56:2.247-259.
- Vovin, Alexander 1993b. *Notes on Some Japanese-Korean Phonetic Correspondences*.  
In Japanese/Korean Linguistics, vol. 3, 338-350.  
Stanford University Press.
- Vovin, Alexander 1994. *Long-distance Relationships, Reconstruction, and the Origins of Japanese*.  
In Diachronica XVI: 95-114.

DOLGOPOLSKY'S THEORY OF STABILITY  
VS. UTO-AZTECAN SECOND PERSON SINGULAR PRONOUNS.

Alexis Manaster Ramer  
Wayne State University

Dolgopolsky (1964) proposed a precise method of arriving at probabilistic classifications of languages, based on a series of claims about the stability of morphemes/lexemes with certain meanings, together with a way of dividing all (consonantal) sounds into classes which could be treated as equivalent for the purposes of establishing the possibility of a formal relationship between morphemes of two or more languages. For example, all labial obstruents are treated as one class, all laryngeals together with an initial velar nasal and zero define another class, and so on. Probable language groupings are arrived at by looking for languages where the stablest meanings are expressed by morphemes made up of identical sequences of these consonantal classes.

This method, which Dolgopolsky used just once (to argue for the plausibility of what has come to be known as the Nostratic language family), deserves to be much better known than it is, and to be both tested and refined. Certainly, it would seem on the face of it superior to the more traditional but less explicit work of, say, Strahlenberg (1730; regarding whose contributions, see Manaster Ramer and Sidwell 1997) or Greenberg (1987), which has, however, received vastly more attention. Indeed, it is one of the more interesting attempts to offer a precise formulation of the intuition that related languages tend to have similar-looking (and not merely regularly correspondent, as some contemporary authors argue) forms for certain particularly basic meanings. This is not to say that we can assume Dolgopolsky's claims to be correct in all details--or even in general. As I said, the method needs to be investigated closely--yet it has not been. As far as I know, there has been no such

work, only unjustified reliance on his hypothesis on the one side and flat disbelief in it on the other.

For the present, I will discuss some rather minor points having to do with second person singular (2sg.) pronouns. In his original paper on this subject, Dolgopolsky (1964) noted fewer than three clear cases of replacement of 2sg. in the languages he had investigated, and took this to be one of the most stable meanings in all languages. Soon thereafter, Dolgopolsky (1965) argued that the attested replacements of 2<sup>nd</sup> (and occasionally 1st) person pronouns was a function of the politeness taboos supposedly found only in languages spoken by relatively complex societies that arose rather recently in historical times, especially in Western Europe and East and South Asia. As a result, such replacements should not be considered relevant to the study of the languages of the simple pre-agricultural societies which we assume for the time scales involved in such controversial proposals for language classification as Pedersen, Illič-Svityč, and Dolgopolsky's Nostratic, Greenberg's Amerind, or the like.

However, the Uto-Aztecan (UA) languages, most of them spoken by small, simple, and sometimes purely hunter-gatherer, societies, seem to offer some counterexamples to this claim. In UA, we find instances of politeness conventions whereby 2sg. pronouns are replaced by other forms synchronically. There are also cases where the 2sg. pronouns in one language appear to be completely unrelated to those in another. If they are indeed unrelated, then we have to assume a diachronic replacement of a 2sg. pronoun. Moreover, even if the pronouns themselves are not replaced, their phonetic degradation over time may be so extreme as to make it difficult or impossible for Dolgopolsky's method to identify the true genealogical connections.

I begin with two examples of politeness conventions in the attested languages.

In Tubatulabal, spoken before the rather recent contact times by a hunter-gatherer population (Voegelin 1938: 11), we find that a son-in-law used the plural to address (in the 2<sup>nd</sup> person) or to refer to (in the 1st or 3rd person) his



mother-in-law, the plural also being used in Tubatulabal myths to express "frigid politeness, scornfulness" (ibid, 44). I do not know how widespread this use of plurals is in "mother-in-law languages" the world over, or whether in any language this leads ultimately to the wholesale replacement of a 2sg. pronoun, but this phenomenon does seem inconsistent with the spirit of Dolgopolsky's (1965) claims.

Another interesting usage is found in Guarijio, which has a special style of speech (called "hablando por dos" or "habla de los compadres"), used among certain in-laws and gossips, in which the 1pl. is used with an active verb form to denote the 1sg. and with a passive verb form to denote the 2sg. While Guarijio is spoken by an agricultural population which has long lived in a contact situation, and hence is not directly a problem for Dolgopolsky's thesis, it seems not unlikely that this kind of usage may have existed in prehistoric times in other UA languages, some of which appear to use reflexes of the UA 1st plural pronouns for the 2sg.

Before proceeding, it should be noted that we will not be interested here in anything that can be deduced about UA pronouns by the use of the method of comparative reconstruction. The reason is simple: Dolgopolsky's proposals concerning stable morphemes were designed to make it possible to arrive at probable groupings of languages which have not been classified definitively before and where, *a fortiori*, there is as yet no reconstruction. In order to test his approach, it is perfectly legitimate then to consider individual UA languages without any reference to Proto-Uto-Aztecan (PUA) reconstructions. What counts is not whether two morphemes are known to come from the same PUA etymon, but only whether their synchronic shapes, when stripped of vowels and then reduced to Dolgopolsky sequences (i.e., sequences of Dolgopolsky's classes of consonants), are identical.

I now turn to a brief discussion of some problems for Dolgopolsky theory, involving some of these synchronic 2sg. forms.

In Nahuatl, we find 2sg. *teh* (independent) and *ti-* (subject prefix), *mic-* (object prefix) and *mo-* (possessive prefix). On the other hand, Cupeño has the following forms: free subject *'e'e* or *e'ep*, free object *e'ey*, enclitic *-e*, *-ep*, or *-p*,

subject/possessive prefix *e-*, object prefix *i-*.

Now, as noted, in Dolgopolsky's way of comparing forms between languages, the only thing that counts is (classes of) consonants in different positions, and so the Cupeño forms would be considered to represent various combinations of the  $\emptyset$  (=zero) class (which includes laryngeal consonants) and the P class. The final *y* of the object form is clearly segmentable as the objective case, and so can be ignored (otherwise, it would belong to the Y (or, as Dolgopolsky writes it, 'J') class).

On the other hand, the Nahuatl forms would be regarded as instantiating the T or the M class, provided we may segment *the* into *t-eh* (as suggested by the 1sg. *n-eh*, 3sg. *y-eh*) and identify the final consonant of *mic* as the object marker, probably cognate with the Cupeño one, (cf. Nahuatl 1sg. *neech-*, which contains the same suffix in a slightly different, and so far unexplained, form). Otherwise, it would count as belonging to Dolgopolsky's K class.

There is thus synchronically no pair of forms from each language that would represent the same Dolgopolsky sequence. If we ask how this situation came about, the answer is complex and by no means clear in all details. The PUA 2sg. pronominal forms have not been adequately reconstructed, although it has long been recognized that there was an M (i.e., \*mV) or  $\emptyset$ -M (i.e., \*Vm or \*VmV) form (see Langacker 1977). Kaufman (1981) also suggests that there may also have been a P (i.e., \*pV) form, which I myself believe was probably used specifically for the subject (in what was therefore a suppletive system).

As for the T forms in Nahuatl, these appear to be an innovation, perhaps formed from the 1pl. stem. The 1pl. pronoun in PUA was a T form, probably simply \*ta. In Nahuatl this yields *tehwaan* (free form), *ti-* (subject prefix), *teech-* (object prefix), and *to-* (possessive prefix). The subject prefixes of 2sg. and 1pl. are actually identical (*ti-*), and the patterning of the independent forms in this language suggests that we can view the pair 2sg. *the* : 1pl. *tehwaan* as analogous to the pair 3sg. *yeh* and 3pl. *yehwaan*. It thus seems likely that the 1pl. is the source of the T forms of the 2sg. If so, then Pre-Nahuatl underwent a partial replacement of its 2sg. It is true, of course,

that the society that spoke the prehistoric stage of Nahuatl at which these replacements would have taken place was presumably already agricultural, so that there is no direct violation of Dolgopolsky's claims, but it is unlikely that it was a society nearly as complex as those in Europe and Asia which alone, according to Dolgopolsky (1965), tend to replace their basic pronouns.

Moreover, Tubatulabal, which as mentioned was spoken by a purely hunter/gatherer population until the rather recent contact times, also has a T form in the 2sg. Specifically, although the subjective case is a P form, *-bi*, the objective case is *-ding*. The existence of the 2sg. possessive *-ing*, and the parallelism with the 2pl. (objective *-dulu*, possessive *-ulu*, should make it clear that the initial d (from \*t) is a separate morpheme, which could in all likelihood come only from the same source suggested for Nahuatl.

I should add that there is another, formally less (but semantically perhaps more) plausible source for the T forms in Nahuatl and/or Tubatulabal: Tarahumara has 2pl. *tumu-he* (independent), *-tu(mu)* (clitic), where the *-mu* element is probably the PUA plural suffix, leaving the *tu-* as a conceivable analogue to the T forms in the other two languages. But this would still mean a partial replacement of the 2sg. in Nahuatl and Tubatulabal.

One could, of course, suggest that this hypothetical T etymon itself reflects a PUA singular (lost in Tarahumara), but the existence of double suppletion: M (i.e., \*mV) or  $\emptyset$ -M (i.e., \*Vm or \*VmV), P (i.e., \*pV), and now also T (i.e., \*tV) in PUA seems like too much of a good thing. Moreover, this kind of proliferation of proto-forms would be a disaster for a statistical method like Dolgopolsky's, for it would drastically increase the chances of chance (i.e., spurious) matches.

As for Cupeño, which was also spoken by a nonagricultural population in pre-contact times, it has traces of the P form, but the loss of final consonants in this language makes it impossible to tell whether some of the synchronic forms do not reflect the hypothetical PUA  $\emptyset$ -M form (or the even more hypothetical T one). In other words, a form like *'e'e* could conceivably reflect something like \*em or even

\*et. It is thus possible that Cupeño did not actually lose the relevant UA etyma, but only reduced them so much that they are no longer identifiable by Dolgopolsky's method. Such purely phonetic problems for his approach are something he does not discuss, although the existence of cases like Russian 1sg. *ya* (which would count as Y) vs. Dutch *ik* (which would represent  $\emptyset$ -K) or English *I* (which would presumably be regarded as just  $\emptyset$ ), even though they all come from the same Indo-European etymon, should perhaps have alerted him to this issue.

In any case, it appears that, if we choose carefully, we can find UA languages whose attested 2sg. pronouns do not exhibit the kind of similarities in terms of Dolgopolsky sequences that his theory would predict. This fact shows that we cannot presuppose an absolute stability of 2<sup>nd</sup> person pronouns in arriving at probabilistic classifications of languages whose prehistories are not known.

I should add that, if we allowed 2pl. forms into the equation, we would be able to find some parallels in terms of Dolgopolsky sequences even between Cupeño forms like *em-* and the like and Nahuatl forms like *am-* and the like, but this kind of extra latitude would, again, increase the likelihood of chance matches, and is in any case not called for by Dolgopolsky.

Of course, it may be that 2<sup>nd</sup> person pronouns TEND to be stable to a very high degree (as proposed by Dolgopolsky 1964), but that is a much weaker claim than the one in Dolgopolsky (1965). Nor is it clear that even the relative degree of stability is really as high as claimed by Dolgopolsky (1964).

Moreover, as Manaster Ramer (1996) points out, a quick look at a number of languages claimed by Greenberg (1987) to belong to Amerind shows that many of them do not have the 2<sup>nd</sup> person \*m (or 1<sup>st</sup> person \*n) pronouns which he postulates as typical of that language family. Of course, the Amerind hypothesis may be false, but it might also be worthwhile considering the possibility that 2<sup>nd</sup> (and perhaps also 1<sup>st</sup>) person pronouns have only been highly stable in certain parts of the world and at certain periods in history. Dolgopolsky's sample was skewed towards the languages of the northern and northeastern parts of the Old World, and it is possible that he

simply missed the relevant examples. There are other classes of meanings which exhibit such non-uniformitarian behaviour, that is, which are stable in certain language families and at certain times only (e.g., 'hand' in Semitic but not in Indo-European, kinship terms in both Semitic and Indo-European but not in UA, numerals in most families of relatively limited time depth but not in, e.g., Afro-Asiatic, etc.). So why not pronouns?

In conclusion, we have shown, first of all, that constructions for expressing politeness or related pragmatic concepts by using something else (1pl. or 2pl.) in place of the regular 2sg. forms are by no means absent from languages of hunter/gatherer and other "simple" societies. Second, there seem to be cases of replacement of 2sg. forms (in certain morphosyntactic slots anyway) by other forms (probably 1pl. or 2pl., again) in (relatively recent) prehistory. Third, purely phonetic processes may alter the shapes of 2sg. forms to the point that their etymological connections become difficult to identify even when using the method of comparative reconstruction on a well-known and close-knit language family, and would be utterly impossible to identify by Dolgopolsky's method applied to languages which are at best very distantly related. We have also suggested that stability (even of pronouns) may vary from place to place and from era to era.

I do not know how seriously these problems would affect the possibility of some revised version of Dolgopolsky's method, but they do seem to suggest the need for a conceptually deeper and factually broader investigation than that undertaken by Dolgopolsky in the sixties. It would be premature to speculate one way or the other about the chances of success for such an investigation, but the topic certainly seems intriguing.

## LITERATURE

- Andrews, J. R. 1975. *Introduction to Classical Nahuatl*.  
Austin & London: University of Texas Press.
- Dolgopolsky, A[ron] B. 1964. *Gipoteza drevnejšego rodstva jazykovyx semej Severnoj Evrazii s verojatnostnoj točki zrenija*.  
In Voprosy jazykoznanija 1964(2): 53-63.
- Dolgopolsky, A[ron] B. 1965. *Soxranjaemosť leksiki, universalii i areal'naja tipologija*.  
In Lingvističeskaja tipologija i vostočnye jazyki, 189-198, Moscow: Nauka.
- Greenberg, Joseph H. 1987. *Language in the Americas*.  
Stanford: Stanford University Press.
- Hill, Jane H. and Rosinda Nolasquez (eds.) 1973. *Mulu'wetam: The First People. Cupeño Oral History and Language*.  
Banning, CA: Malki Museum Press.
- Kaufman, Terrence. 1981. *Comparative Uto-Aztecan Phonology*. Ms.
- Langacker, Ronald W. 1977. *An overview of Uto-Aztecan grammar*.  
In Studies in Uto-Aztecan Grammar (ed. by Ronald W. Langacker), v. 1.  
Dallas: Summer Institute of Linguistics and the University of Texas at Arlington.  
[= Summer Institute of Linguistics Publications in Linguistics, 56.1.]
- Lionnet, Andrés. 1972. *Los elementos de la lengua tarahumara*.  
Mexico City: Universidad Nacional Autónoma de México.
- Manaster Ramer, Alexis. 1996. *Tonkawa and Zuni: Two case studies on the validity of the Greenberg classification*.  
In International Journal of American Linguistics 62:264-288.
- Manaster Ramer, Alexis; and Paul Sidwell. 1997. *The truth about Strahlenberg's classification of the languages of northeastern Eurasia*.  
In Journal de la Société Finno-ougrienne 87:139-160.
- Ševoroškin, Vitalij V. and Markey, Thomas L. (eds. and trs.) 1986. *Typology, Relationship, and Time*.  
Ann Arbor: Karoma.
- Strahlenberg, Philipp Johann von. 1730. *Das nord- und ostliche Theil von Europa und Asia. ...*  
Stockholm: In Verlegung des Auctoris. Reprinted (1975), with an introduction by J. R. Krueger. Szeged: Universitas Szegediensis de Attila József Nominata. [= Studia Uralo-Altaica 8.]

Voegelin, Charles F. 1935a. *Tübatulabal grammar*.

In University of California Publications in American Archaeology and Ethnology 34:  
55-190.

Voegelin, Charles F. 1935b. *Tübatulabal texts*.

In University of California Publications in American Archaeology and Ethnology 34:  
191-246.

Voegelin, Erminie W. 1938. *Tübatulabal ethnography*.

In Anthropological Records 2.1: 1-90.





# THE PRESERVATION OF NOSTRATIC WORD MEANINGS AND SOUNDS, STATISTICAL DATA

James Parkinson

From the yet-to-be constructed original language, one second-generation proto-language is arbitrarily called Nostratic (Latin, meaning "our language"). Encompassing Europe, North Africa, and most of Northern and Western Asia, the Nostratic languages are spoken by a bare majority of the world's population today. Of the six third-generation proto-languages that comprise the Nostratic group, a preliminary analysis of Illič-Svityč's Nostratic Dictionary and Dybo's Comparative Phonetic Tables suggests the approximate degree to which each has preserved Nostratic meanings and sounds:

<u>PROTO-LANGUAGE</u>	<u>MEANINGS, %</u>	<u>PHONETICS, %</u>
Semito-Hamitic	44	69
Uralic	41	66
Altaic	44	55
Indo-European	47	40
Kartvelian	23	76
Dravidian	34	42

Semito-Hamitic (or Afro-Asiatic; surviving in Hebrew, Arabic, Ancient Egyptian, Berber, Chadic and Cushite languages) experiences a majority of its phonetic changes in the affricate sounds: *z* and *c* ( $\approx$  *ts*) in their various forms. Otherwise, Semito-Hamitic has changed phonetics no more than Kartvelian. (Hebrew and Egyptian apparently became mutually unintelligible between the 21<sup>st</sup> and 19<sup>th</sup> centuries BC; Ge 12:10-19 and 42:23. By the 13<sup>th</sup> century BC dialectical differences had become apparent in Hebrew; Jg 12:6.)

Uralic (surviving in Finnish, Estonian, Hungarian, and the Samoyed languages of the north Russian tundra) poorly preserves postvelar, and pharyngeal and laryngeal sounds: **q** and **h** in their various forms. Otherwise, it does well.

Altaic (surviving in Turkic languages, including Kazakh and Turkish, Mongolian, the Tungus languages of northern Siberia and Manchuria, and probably Korean; found in Turkey, but mostly east of the Caspian Sea and east of the Yenesei River) preserves the affricate sounds as well or better than most.

Indo-European (surviving in English and other Germanic languages, Celtic and Slavic languages, Latin, Greek, Tocharian [= Togarmah] in central Asia, Iranian languages, Sanskrit and Hindi; encompassing two thousand million native speakers, or over 40% of the world's population in Europe, Iran, India, and now the Americas and South Pacific) preserves sonant and sibilant sounds (**r**, **l**, **m**, **n**, **j**, **w**; **s**) better than most other sounds. Meanings of words are relatively well preserved.

Kartvelian (surviving primarily in Gruzian [= Georgian], the modern language of Tubal, centered in Tbilisi and the Kura River valley; also Mingrelian, Svan, Chan or [unwritten] Laz) preserves pronunciations better than most, though apparently it has totally lost more words than the other proto-languages. Meanings of words are not relatively well preserved.

Dravidian (surviving in Tamil, Kannada, Malayalam, and to a lesser extent in Telugu, mostly in South India) preserves affricates and sibilants - **c**, **z**, and **s** in their various forms - better than most other consonants, and especially better than combined consonants such as **-Sb-**, **-Sd-**, or **-Lg-** (where S and L each represent several similar sounds).

Table I

## Preservation of Nostratic Meanings and Phonetics

Nostratic	Meanings	Phonetics		
<u>Proto-Language</u>	<u>from</u> <u>Nostratic</u>	<u>"A" from</u> <u>Nostratic</u>	<u>"B" from Proto-</u> <u>Language</u>	<u>"A X B"</u>
<b>Kartvelian</b>	23.4%	76.2%		
Georgian (Gruzian)			85.5%	65.2%
Zan (Mingrelian & Chan)			69.2	52.7
Svan			76.1	58.0
<b>Semito-Hamitic (Afro-Asiatic)</b>	44.4	69.3		
Semitic			97.8	67.8
Ancient Egyptian			67.2	46.6
Berber			67.2	46.6
Cushite			73.9	51.2
Chadic			75.6	52.4
<b>Uralic</b>	40.7	66.4		
Balto-Finnic			73.1	48.6
Mordovian			≈64.9	≈43.1
Ob' Ugric			≈64.8	≈43.0
Lapp (Saam)			≈60.9	≈40.4
Mari			≈60.2	≈40.0
Samoyedic			≈48.4	≈32.1
Permian			≈46.9	≈31.1
Hungarian			≈42.9	≈28.5
<b>Altaic</b>	44 ±3	55.4		
Turkic			79.2	43.9
Mongolian			74.7	41.4
Tungus			84.6	46.9

<b>Dravidian</b>	33.8	41.6		
Tamil			87.5	36.4
Kannada (Canarese)			86.0	35.8
Telugu			85.6	35.6
Gondi			68.9	28.7
Brahui			68.9	28.7
<b>Indo-European</b>	47.5	40.05		
Indo-Iranian			61.8	24.7
Anatolian			55.3	22.1
Armenian			52.0	20.8
Greek			67.1	26.9
Italic			71.1	28.5
Germanic			61.8	24.7
Balto-Slavic			66.7	26.7

The initial list has been expanded to include the phonetic changes of several language families, as shown in Table I. The preservation of Nostratic meanings has been evaluated for the proto-languages only and is included in Table I. The method for compiling the percentages is sketched below; it is not above improvement.

#### Phonetic Preservation from Nostratic to its Proto-Languages

The preservations of Nostratic phonetics are evaluated from Dybo's Comparative Phonetic Tables. The six proto-languages are effectively evaluated on a scale from 0 (poor preservation) to 4 (fully preserved) for Nostratic consonants.

For examples, each of five proto-languages receives a score of 4 for preserving **t-**, while Indo-European (I-E) receives 0 for shifting it to **d**. Partial credit is given for minor phonetic shifts, as a shift from **p** to **b** (2 points), or a shift from **λ** to **l** (3 points). The stops and sonorants are given 0, 2, or 4, while the fricatives, affricates, pharyngeals, etc. are given 0, 2, 3 or 4. Where the correspondences were not known in Dybo's tables, they are overlooked. Thus, the overall preservation is calculated for each proto-language by adding up all scores and dividing by 4 times the number of

known correspondences. (It is suspected that simply ignoring the unknown correspondences might slightly bias a proto-language, such as Dravidian, to the up side, but overall it still shares with I-E the distinction of preserving Nostratic phonetics most poorly). Differences of only a few percent are likely not significant, although differences of 5-10% likely are.

#### Phonetic Preservation from Proto-Languages to Daughter Languages

Preservation of the phonetic correspondences from each proto-language is evaluated for each sound on a scale from 0 to 3, with resolutions of either  $\frac{1}{2}$  or 1. (Vowels are not included for Semito-Hamitic and Indo-European.) The procedure is otherwise the same as above.

Semitic is notable for its high preservation of Semito-Hamitic consonantal sounds (the vowels have not been reconstructed). By contrast, the preservation of Uralic sounds in the Hungarian, Permian, and Samoyed groups is unusually poor (possibly reflecting their long migrations, BC 1500 to AD 500).

#### Preservation of Meanings from Nostratic to its Proto-Languages

The preservations of Nostratic word meanings are evaluated from Mark Kaiser's translation of the headers in the three current volumes of Illič-Svityč's Nostratic Dictionary. Each word is evaluated on a scale from 0 to 5, from minimal association to full preservation of the meaning. (There are few scores of 1 or 0, as the reconstruction itself would usually have remained in doubt, rendering it an unlikely entry in the Nostratic Dictionary.)

There is no particular correspondence between preservation of phonetics and preservation of meanings: Kartvelian preserves phonetics better than the others, but alters the meanings most. Indo-European preserves meanings as well or better than others, but poorly preserves the phonetic sounds. Overall, Nostratic is apparently best preserved in Semito-Hamitic and probably Uralic.

## LITERATURE

- Dybo, Vladimir, 1989, 1990. *Comparative Phonetic Tables*.  
in BPX 23, p. 114-121, repeated in BPX 25, p. 168-175  
[For complete tables, see *Nostratic Dictionary*, Vol. I, below].
- Illich-Svityč, V.M. 1967, *Materialy k sravnitel'nomu slovarju nostratičeskix jazykov*. [Materials for a comparative dictionary of the Nostratic Languages],  
in Etimologija 1965, p. 307-373 [Russian text and meanings]
- Illich-Svityč, V.M. 1971, 1976, 1984, *Opyt sravnenija nostratičeskix jazykov: Sravnitel'nyj slovar'* [A Comparison of Nostratic Languages, Comparative Dictionary], 3 vols. to date,  
Nauka, Moscow [Commonly referred as the *Nostratic Dictionary*, 378 nostratic words, russian text and meanings, with indexing in Russian and English. 3 words with full translation by Mark Kaiser are given in BPX 23, p. 122-127].
- The Nostratic Reconstructions of Illich-Svitych*  
Translated by Mark Kaiser, in BPX 25, p. 138-167
- Kaiser, Mark, 1990, *Semantic Index to Nostratic Reconstructions*  
in BPX 25, p. 176-197. See also:
- Parkinson, J., 1989, *A Nostratic Word List: Reconstructions by V. Illich-Svitych*  
in BPX 23, p. 138-167 [688 independent Nostratic words].
- Materials from the First International Interdisciplinary Symposium on Language and Prehistory, Ann Arbor, 8-12 Nov. 1988; ed. Vitaly Ševoroškin, Bochum, Universitätsverlag Dr. Norbert Brockmeyer (BPX series, 5 vols.):
- Reconstructing Languages and Cultures*, BPX 20, 1989.  
*Explorations in Language Macrofamilies*, BPX 23, 1989.  
*Proto-Languages and Proto-Cultures*, BPX 25, 1990.  
*Dene-Sino-Caucasian Languages*, BPX 32, 1991.  
*Nostratic, Dene-Caucasian, Austric and Amerind*, BPX 33, 1993 (with index).

COMMENTS ON SERGEI STAROSTIN'S PAPER  
ON LINGUISTIC DATING

Henrik Birnbaum

I can certainly understand that it would be very desirable for us to have a reliable tool for linguistic dating and, more specifically, decay dating, especially when concerned with distant relationships and “deep” reconstruction of protolanguages, or rather, what I have called “preprotolanguages”. I also agree that the glottochronological or lexicostatistic technique of Morris Swadesh and his followers (claiming a constant rate of lexical turnover, or replacement, of 14 per 100 lexical items over 1000 years), proved highly unreliable, not to say faulty. I have therefor also expressed my skepticism concerning this (original) kind of glottochronology in my book (1977) on linguistic reconstruction (pp. 17-18), which cites critical work by R. Anttila and I. Fodor. Thus, I can say that I share the general rejection of Swadesh’s glottochronology as a viable tool for establishing absolute chronologies in linguistic reconstructions, especially as applied to the distant past and to dating what is conceived as language “splits”. In fact, I consider this overly simplified method no more revealing than W. Manczak’s more recent attempts to localize ethnolinguistic proto-homes (of the Goths within the Germania; the Slavs in relation to Balts, Germanic tribes and Iranians; or, for some matter, the original Indo-Europeans) solely on the basis of “measuring” the degree of genetic relationship among languages by lexical counts (applied to chronologically disparate texts) and projected onto spatial relations. I would readily agree that the improved – “calibrated” – method proposed by Dr. Starostin, represents genuine progress. Thus, it is no doubt a substantial improvement to exclude loanwords from any lexical count and to restrict it to root etyma. Therefor, remaining this technique of “root chronology” or “etymostatistics” and interpreting the mathematical formula differently, viz., in transcendental, rather than exponential terms. Yet, I am still

First, I would like to begin with a few words about Scandinavian and Slavic data, with which I am familiar. Here, I must say that it was not, perhaps, particularly fair of the Norwegian linguists (K. Bergsland and H. Vogt) to test Swadesh's method by contrasting contemporary Norwegian (in the *bokmal*, formerly *riksmal*, variant, by Hauger labelled "Dano-Norwegian") with modern Icelandic data. For both these languages represent, in a sense, extreme, indeed exceptional instances. The Norwegian standard language (contrary to New Norwegian, or *nynorsk*, formerly known as *landsmål*) does not, in fact, represent a naturally grown language. Rather, it must be considered a hybrid formation of sorts: a historically conditioned blend of genuine Norwegian (i.e., in genetic terms, West Scandinavian) phonology (which shares some basic phonological characteristics – e.g., the treatment of velars before front vowels and the distinction of two types of pitch – with genetically East Scandinavian Swedish!) And essentially Danish-type, i.e., historically East Scandinavian, morphosyntactic structure. In vocabulary, the Danish elements cannot simply be considered loanwords (and thus eliminated in Starostin's model), but rather constitute a whole, substantial lexical layer due to a historically conditioned (and well understood) Danish superstratum. In this context, I would also like to recall that according to an uncontested count made by E. Wessén, contemporary standard Swedish contains more lexical items from Low German origin, than such that can be traced back to Proto-Nordic (Common Scandinavian). There is also the example of modern-day Hungarian, wherein the share of genuine Ugri lexical items, though of high frequency, is quite small as compared to words of foreign provenance. Or, how about modern English? Are we really to eliminate all Romance (Old French), Latin, Greek, and even Danish loans, which have been entered the language in the course of the last 1000 years or more – in other words, are we to consider only the pure Anglo-Saxon lexical stock (minus even any Danish mixture) when applying "etymostatistics", as now defined, to modern English? Does analogous consideration apply to languages such as standard Norwegian, Swedish or Hungarian? What about modern Icelandic? As we know, contemporary Icelandic is a particularly conservative



language (and incidentally, that is why today's Icelander can, without much difficulty, read the sagas and Edic and skaldic poems of his ancestors). However, modern Icelandic is highly archaic, not only because of its isolated, insular position, but also due to a deliberate, consistent, and sustained purist language policy effectively coining new words to denote – particularly technical – concepts by means of inherited lexical material. This is comparable to the purist wave (of relatively short duration, to be sure), that gave modern German such words as *Bürgersteig* (along with *Trottoir*), *Fernsprecher* (along with *Telefon*), *Abort* (along with *Toilette*), and the like.

As for Slavic (and, *mutatis mutandis*, similar considerations apply to other language groups), what exactly do we mean when claiming that, on the basis of lexical counts of (for example) Russian and Polish, we can date the “split” of Common Slavic into several branches (presumably, the traditional East, West and South Slavic groups) or possibly even individual languages to the beginning of our era? As is well known, there exist substantial – and, I would add, justified – differences of opinion as to how to date the disintegration and eventual dissolution of the Common Slavic parent language. (This is how I would like to refer to it, given that “split” suggests a sudden event or an evolution of at most short duration.) Depending on whether the focus is on the earliest ascertainable innovations (which did not embrace all of Slavic), or whether we prefer to pay attention to the last retentions shared by Slavic as a whole, we may date the final period of Common Slavic anywhere between the 5<sup>th</sup> and the 12<sup>th</sup> century A.D. (But, hardly as early as the 1<sup>st</sup> B.C. or A.C.). For example, the so-called 2<sup>nd</sup> regressive and the progressive (Baudouin) palatalization did, as every slavist knows, not affect the Slavic linguistic territory in its entirety. Moreover, some sound shifts whose reflexes can be found everywhere in Slavic today, may initially have been more restricted, even though their earlier dialectal nature can no longer be recovered.

A lexicostatistic interpretation of the degree of relatedness among the Slavic languages (but, not the dating of the disintegration of Common Slavic) was recently

sketched by A.F. Žuravlev in *Vaprosy jazykoznanija* 1988/4: 37-51 (*“Leksikostatičeskaja ocenka genetičkoj blizosti slavjanskix jazykov”*). Though the aim of Žuravlev’s inquiry is different from that of Swadesh and Starostin, whose primary goal was/is to develop a method for establishing absolute chronologies of language splits, much of the critique of Swadesh’s technique voiced in his article can also be levelled against Starostin’s improved “etymostatistics”. This is true, among other things, for the arbitrariness and allegedly diagnostic selection of a highly restricted word list, usually only 100 or 200 items; as well as the total disregard for the phenomenon of secondary lexical interpretation of such closely related languages as the Slavic. However, surprisingly, Žuravlev does not even mention Starostin’s relevant work. While I find Žuravlev’s argument generally noteworthy, this does not, of course, mean that I necessarily agree with everything he has to say.

And I can only agree, obviously, with what was said here by Professor Pulgram with reference to setting a precise (or even approximate) date, or narrowly defined period, when Late (Vulgar) Latin (or Common Romance) was no longer Latin (or Romance) but was already Spanish, French, or Italian. For, we are dealing here with an unbroken continuum in time (and, at least early on, also space). The decision to consider one phase still (Late) Latin and as a subsequent stage already (Early) Castilian (but, hardly yet Spanish), (Old) French, or (Early) Italian (Tuscan, etc.) is entirely arbitrary. In other words, it is dependent on the criteria – and lexical are certainly only one kind, with grammatical and/or phonological claiming equal legitimacy – we choose to select or focus on. Of course, the very concept of language is also questionable, since the criteria for its definition – mutual intelligibility being only one among several conceivable – remain controversial or poorly understood and, at any rate, not generally agreed upon. Perhaps, we should therefor really not speak of “languages” and “language splits”, but merely, as Professor Lamb has suggested, of shared “sets of inherited etyma”, which is a lot less than “languages”, since the latter, however defined, always has the two facets – the expression and the component plane.

It is ironic, in the sense that it is for those thoroughly studied and etymologically largely transparent language families and relatively short spans (of one or a few millennia), where our need for reliable dating techniques is felt the least. The improved method brought to our attention can and must be tested and – I am afraid – still fails. Particularly, I find the insistence on a constant, cultural-context-free value – now, after eliminating all borrows, allegedly 4 or 5 rather than 14 per 100 over one millennium – entirely unacceptable since, obviously, we must still reckon with a great number of variables affecting the rate of linguistic, or even root-morphemic, change. For such changes can be broadly defined as culturally and/or socially conditioned.

I am also bothered by the criteria (or lack thereof) for selecting the corpus of the basic 100 lexical items. How is it arrived at? On the basis of frequency? Or by using some sort of basic-concept list? However, if the latter is the case, how do we determine which first 100 (or even 200) items are more basic than the next 200 (or 300, 400, etc.)? The “basic vocabulary” of particular languages will surely contain more than just 100 or even few hundred items, and moreover, will greatly vary from language to language, and from culture to culture, and from one degree of a society’s sophistication to another degree, and so forth. And if for Russian (or other Slavic languages) we only have to include two terms for bodily extremities, R *naga* and *rufa*, but for the most other Indo-European languages four (say, E *leg*, *foot*, *arm*, *hand*) by which added two items can we fill the Slavic “vacancies”? Or will R *palec* and both E *toe* and *finger* (again with the same asymmetry) be part of the basic 100-word list? And, if so, why, and at the expense of what other perhaps no less “basic” lexical item? And while ‘water’ (or perhaps, a more generalized concept for ‘liquid’ or ‘fluid’) presumably will be in every speech community’s list. How about words for ‘ice’, (or even ‘snow’) – central for Eskimos and Scandinavian, but hardly for Arabs or the people of Southeast Asia? In the same vein, if indeed we would have to construct different 100 or 200-word lists for different speech communities or cultures, the whole point, the very base for any such computation would be lost, or

would it not?

And one more thing: should our linguistic dating, even for prehistoric evolutionary stages, really be concerned with divergent branching and language “splits” (as represented by the diagrammatic tree) and not also with convergent merging, language mixing, and creolization? As I have tried to suggest in my own paper, even if we accept the notion of a Nostratic macro-family or even Proto-Nostratic, must we not realistically assume that their prehistory in turn combines process of divergence, as well as convergence? Is it not, somehow, absurd to test the proposed method on, for example, contemporary standard Russian, when we know perfectly well that modern Russian itself is the product of two Slavic languages, namely, a variety of East Slavic vernacular, and essentially South Slavic Old Church Slavonic?

Therefore, while acknowledging the significant improvement of Dr. Starostin’s root-chronological, or etymostatistical, technique over Swadesh’s original method, I still cannot find the new “calibrated” statistical method for linguistic dating a genuinely reliable tool for the purpose for which it was devised.

Part II:  
Genetic Relationship of Languages  
And "Mass Comparison".



PHYLUMPHILE OR PHYLUMFOE?  
REEXAMINING GREENBERG'S METHOD OF MASS COMPARISON

Naomi Gurevich

1. INTRODUCTION

Classification of languages is an undertaking that involves the comparison of a large number of lexical data. An attempt to trace genetic relationships between languages and to reconstruct language families, or phyla, that engulf a great number of languages spoken today is fraught with problems and controversy. Most significant of these is the deficiency in data: The greater the time depth of the data examined, the smaller the attested data that can be useful for this work. It is this difficulty that Greenberg attempts to transcend by quantifying the amount of data and the degree of relationship necessary for the classification of languages. To this end Greenberg proposes an approach, termed Mass Comparison, which involves comparing a limited amount of words across many languages instead of the more traditional strategy of comparing a large number of words across a few languages at a time.

Greenberg's claim that simultaneous comparison of many languages can provide very accurate criteria by which to classify the languages in question is the central issue discussed in the present paper. It is important to note, however, that the number of words to use in such comparisons is also a matter of controversy. Of the more significant lists available for comparative work is the list of 15 most

stable words, ordered by stability (Dolgopolsky, 1964), as well as attempts to continue this ordered list. Work on word lists is also carried out by Starostin, whose paper on Lexical Statistics in this volume discusses the method of "Root Glottochronology" and its ability to transcend some of the difficulties of word lists, such as choosing a basic word from a list of synonyms. All these are vital issues within comparative work—whether it uses Greenberg's Mass Comparison method as discussed in the present paper, comparative methods as traditionally practiced, or any methods in between. However, the question of which word lists can best serve comparative methods is not central to the present paper which aims to continue a study initiated by Peter (1991).

Peter's study launches an empirical test of Greenberg's method of Mass Comparison by comparing 83 basic notions, a trimmed down version of Swadesh's (1967) original basic vocabulary list, in three languages whose history is known: English, Hindi, and Finnish. The vocabulary from the three languages is first "eyeballed" for evidence of similarities; then etymology is consulted to determine the possible origin of any similarities detected, and the ratio of total similarities to those not indicative of a genetic relationship (i.e. similarities resulting from borrowing, chance, onomatopoeia or nursery language) is calculated. Peter concludes that Greenberg's method is not productive: At best the method had only a 50-50 chance of predicting genetic relationship. 66% of similarities found were, upon investigation, judged to be "false friends". The number of lexical items that resembled each other in all three languages was originally measured at 11.9%, but none of these was shown to reflect genetic relationship between the three languages.

While Peter's study calls the value of Mass Comparison into question, further investigation is needed to address Greenberg's (1963, in Ringe 1995, p.62) claim that the larger the number of languages examined, the greater the accuracy of the method in predicting genetic relationship between them. To this end additional languages are added to Peter's original data: Hock (work in progress)



has added German and Marathi, and the present paper adds Russian, to bring the total number of languages compared to 6. The methods outlined in Peter's study are followed as closely as possible, and the ratio of resemblances indicative of genetic relationship to the number of total resemblances is examined. Greenberg's method of Mass Comparison is empirically tested on two levels: 1) Will the number of resemblances between the lexical items of Russian and the other 5 languages be indicative of the distance known to exist between them - i.e. will more Russian words correspond to other Indo-European (IE.) languages than to Finnish? And 2) Will the increase in the number of languages compared also increase the accuracy of the method?

The present study also raises some questions about the methodology of Mass Comparison or, more precisely, the lack of clear guidelines for such work. While Greenberg claims that his approach is powerful enough to circumvent the need to trace sound correspondences, as in traditional reconstruction, most researchers do not hesitate to include sound correspondences in their work, and Peter is no exception. The present analysis, however, follows Greenberg more closely by relying on resemblances between languages that are obvious to the naked eye, without the benefit of tracking sound correspondences. The resulting analysis involves many fuzzy distinctions and arbitrary choices, all of which are discussed in section 3, titled "Methods of Analysis and Data Collection".

## 2. THEORETICALLY-BASED CRITICISM OF THE MASS COMPARISON METHOD

Peter and Hock work to empirically test Greenberg's method of Mass Comparison. In contrast, Ringe (1992; 1995) attempts to discredit this method with the use of statistics and mock data. Ringe concludes that the "Mass" in "Mass Comparison" presents the greatest flaw in this method: Not only does

increasing the number of languages compared fail to increase accuracy, it makes the method less reliable by providing a greater probability for chance resemblances in the data. Ringe's argument is briefly discussed below; a closer examination of his work with mock data, statistics, and the laws of probability will be the subject of a subsequent paper.

Ringe's central claim, that each additional language increases the probability of chance resemblance occurring somewhere in the data, seems mathematically valid. Consider the following hypothetical cases which illustrate Ringe's argument:

Example I	
A	B

Example II		
A	B	C

Example I represents a comparison of lexical items from two languages; example II represents a comparison of lexical items from three languages. In scoring resemblances between the lexical items, as required in the Mass Comparison method, Ringe itemizes the possible combinations of matching words: In Example I only one combination is possible: A-B; in Example II four combinations are possible: A-B, A-C, B-C, and A-B-C. Adding a fourth language would bring the number of possible combinations to 11. If one considers each combination as representing the possible matches between the words compared, it is obvious that more possibilities become available merely by the addition of another word. On the basis of this, Ringe states that as the number of possible combinations of corresponding words increases with each additional language, so does the probability that some such correspondence will occur by chance alone.

Ringe's argument indeed makes good mathematical sense. More potent is the fact that each additional language may add more than just one lexical item to

the list of data, and often as many as 7 synonyms of a particular basic notion<sup>1</sup>. However, Ringe's argument, as demonstrated above, makes no distinction between the methods of Mass Comparison and more traditional methods involving the comparison of large numbers of words across a few languages at a time. In effect, Ringe argues that the greater the number of total items compared, the greater the possibility for chance correspondence. A long word list with only two languages would also, in theory, provide ample opportunity for chance similarities; the larger the list, the greater the probability for such correspondences. And yet it is generally accepted that the longer the word list, the more accurate the comparative work. Obviously, arguments that do not take into consideration actual linguistic factors cannot account for what happens to real language, and are therefore not productive. It is for this reason that an empirical study of the Mass Comparison method is warranted.

### 3. METHODS OF ANALYSIS AND DATA COLLECTION

As mentioned above, the present paper follows the methodology of Peter's original study. The list of 83 notions that allow a wide margin of semantic differences is translated into Russian (Appendix I). Each lexical item in the Russian vocabulary is then compared to its lexical counterparts in the 5 languages: English, German, Hindi, Marathi, and Finnish. A second list is created where only those Russian words that are found to resemble at least one item from another language, along with the corresponding items, are entered (Appendix II).

The process of determining which words resemble each other is by no means consistent. Familiarity with IE. sound changes can be both helpful and deterring: On the one hand, it could help indicate if a particular resemblance between two words is possible according to how closely the similar segments in

---

<sup>1</sup> This same difficulty—choosing a basic word from a list of synonyms—is addressed in Starostin's

these words follow attested sound changes. On the other hand, many other changes—some unrelated to regular sound changes, such as by reasons of taboo—may have influenced the history of a particular word. In such cases expecting specific results that follow Grimm's law, for example, may cloud one's judgment. While these problems and others, discussed below, testify to the need for clearer guidelines, they also combine to illustrate how mock data, as used by Ringe, cannot sufficiently represent real language data and all the factors, linguistic or otherwise, that affect it.

The final step of the study rests in consulting the etymology of all the lexical items found to resemble each other. Words that came into a language as borrowings, nursery language, or onomatopoeia are eliminated from the list; any resemblance these words exhibit to items in other languages is not due to genetic relationship. Similarities between words that are inherited, but originated from different roots or IE. bases, are also discounted as chance resemblance. Finally, only correspondences that represent true genetic relationship are counted<sup>2</sup>, and their number is calculated in terms of its ratio to the total resemblances found in the data. If Greenberg's claim that additional languages increase the accuracy of his method is to be supported, the new ratio, calculated upon comparing 6 languages, should improve in comparison with Peter's results that involved 3 languages.

---

aforementioned paper on Lexical Statistics.

<sup>2</sup> Eliminating all but genetically related matches is the ultimate goal at this stage of the analysis.

However, a limited access to etymological data will keep the present study from reaching this goal at this time.

### 3.1. TO RESEMBLE, OR NOT TO RESEMBLE... THAT IS THE QUESTION

Although this process is central to Mass Comparison, there are no clear criteria for judging when two words can be considered as similar. In one case the words can share many features, such as in the case of the genetically related Russian *nos* and English *nose* (the two words share word-initial nasals, back round vowels, and word-final sibilants); in another, the correspondences are less obvious, as in the case of genetically related Russian *doc'* and Marathi *dh ṭda* (which share only the word-initial dental). Furthermore, near perfect similarities are almost certain to end up as “false friends” where at least one of the words enters its language as a borrowing, as with English *lamp* and Russian *lampa*, or English *fruit* and Russian *frukt*. Three other problems are encountered when comparing lexical items:

The first problem involves cases where a word from one language is thought to resemble more than one word in another language. It is clear that a lexical item is unlikely to be related to two different and unrelated words; it is not often that a word that originated from two different roots is uncovered. What isn't clear is which of the two choices bears greater similarity to the item in question, and how to score such a case.

For example:

English	Russian
irate	serdityj
furious	yarost'

While *furious* resembles *yarost'*, *irate* seems equally likely to be related to *serdityj* (on the basis of the high vowel + liquid /r/ + dental segment) as it is to *yarost'* (on the basis of /y/~/i/ and liquid /r/ segments). After some deliberation *serdityj* is chosen as *irate's* match, when in fact *yarost'* is the true genetic relative of the English word *irate*, which itself turns out to be a borrowing. In counting individual resemblances, especially for the purposes of later calculating the ratio of genetically related matches to total matches, such a case is difficult to score and involves a seemingly arbitrary judgment call on the part of the researcher.

A second problem in scoring resemblances involves a "chain reaction" of similarities, as in the following case:

English	German	Hindi	Marathi	Finnish	Russian
cart	karren	kar	kar	koura	ruka

At first *cart* and *ruka* are not considered a match; but *koura* is found to resemble the Russian word *ruka* (on the basis of corresponding vowels, and consonantal segments that may have undergone metathesis). The rest of the words (*kar*, *kar*, *karren*, and *cart*) are subsequently included in the list of matches due to their similarity to *koura*. The correspondences between Russian and other languages in this particular example, as it turns out, are not indicative of a genetic relationship.

Another example of a "chain reaction" of similarities between languages can be seen in German *floh*, Russian *bloxa* and English *flea*: *floh* matches *bloxa*, and *flea* resembles *floh*, which results in an apparent match between *flea* and *bloxa*. Again, "false friends" are created by this process. However, the strategy of scoring

resemblances by “chain reaction” is also productive in at least one case, where Russian *z'evat'* is thought to resemble English *chew*, which in turn corresponds to German *kaunen*. Hence the genetic relationship between Russian *z'evat'* and German *kaunen* is found where otherwise it would not have been uncovered since the two words seem wholly dissimilar.

A third problem surfaces at the stage when etymology is consulted, but is relevant to this stage of comparison. In some such cases a word that was not originally included in the list of data, but could have been due to its synonymous meaning, is discovered after the stage of comparison. Significantly, a more timely inclusion of the newly found term may have prevented a match of “false friends”. For example, the English word *seek* is matched with the Russian word *iskat'* and found to be of a different origin; but the English word genetically related to *iskat'*, namely *ask*, was not originally provided as a choice. Similarly, Russian *ryz'yy* is mistakenly thought related to English *ruddy*. Had the word *rust* been included under the lexical notion of “red”, it would surely have been picked as more closely resembling *ryz'yy*, resulting in a genetically motivated similarity instead of the “false friends” that *ryz'yy* and *ruddy* turned out to be.

In a similar situation, English *circular* and Russian *kruglyj* are scored as a match. In fact the Russian word, although unrelated to *circular*, is related to *ring*. *ring* was not originally included in the data, but seems appropriate for the lexical notion in question, that of “round”. However, if *ring* had been included, it is doubtful that it would have been chosen as a match to *kruglyj* with *circular* as a seemingly more worthy choice due to the shared features such as word-initial velar, high vowel, liquid /r/, and another velar followed by liquid /l/ (although some familiarity with Russian morphology would rule out the /l/ as a basis for similarity). Finally, the case of *spit*, *saliva*, and the Russian word *plevat'* echoes this same conflict: *spit* is included in the data, but *plevat'* is instead matched with *saliva*; as it happens, *spit* and *plevat'* are in fact related while *saliva* and *plevat'* are not. In effect, the method of Mass Comparison failed to predict the genetic

relationship between *spit* and *plevat'*, and would have just as easily failed to predict the relationship between *ring* and *kruglyj*.

Although the greater goal of Mass Comparison is to predict relationship between languages, not to find each and every genetically related word, the problems discussed above are significant. If there is no fixed manner in which to score resemblances between words, especially without the benefit of tracing sound changes throughout each language, any study that utilizes Mass Comparison becomes difficult to replicate. It is generally accepted that if the methodology of an experiment cannot be replicated, more often than not, the methodology is not reliable.

### 3.2. ANALYZING RESULTS

#### 3.2.1. STAGE I: BEFORE ETYMOLOGY

In the first stage of the study a list is compiled of all the Russian words that are judged to resemble synonymous words from any of the other languages (Appendix II), regardless of the origin of the observed resemblances. If the method of Mass Comparison were to be used on languages whose history is not known, or on reconstructed—not yet attested—languages, such a list of total resemblances would comprise the most significant chunk of data that would be used for either classification or reconstruction. The following step, that of consulting the etymology of the languages used in the present study and weeding out those resemblances that are not indicative of a genetic relationship, would not be available for such work.

A total of 81 Russian terms are found to resemble words with corresponding meaning from a different language. Some of the 81 terms are synonyms; in fact, the 81 lexical items are spread out over only 57 different notions. This is a significant source of possible inconsistencies. One may choose to limit matches to only one per lexical notion, in this way weeding out possible chance resemblances



# PHYLUMPHILE OR PHYLUMFOE ?

at the price of neglecting a few genetically motivated similarities. Another may score every similarity regardless of the fact that more than one inherited term for every lexical notion is unlikely. A similar judgment call is required when counting the matches (or "hits"): Should each hit between lexical items count, or should only hits between lexical notions count? Both perspectives are illustrated in the following two tables:

Table I: Results that count resemblances between lexical items

Total Russian to any language	Russian to English	Russian to German	Russian to Hindi	Russian to Marathi	Russian to Finnish
81	49	42	35	36	30

Resemblances between all 6 languages: 14

Resemblances between all, and only, IE. languages: 7

Resemblances between all languages, except for one IE. language: 2

Table II: Results that count resemblances between lexical notions

Total Russian to any language	Russian to English	Russian to German	Russian to Hindi	Russian to Marathi	Russian to Finnish
57	43	38	29	31	28 <sup>(A)</sup>

Resemblances between all 6 languages: 17<sup>(B)</sup>

Resemblances between all, and only, IE. languages: 7

Resemblances between all languages, except for one IE. language: 0

Although overall the numbers in Table II are lower, the following two examples illustrate that counting only resemblances between lexical notions, in effect ignoring more than one resemblance between various synonyms within these notions, does not always have a limiting effect on the results: <sup>(A)</sup> Lexical items #43 and 60 in Finnish (*mies*) and Russian (*muz'*, *muz's'c'ina*) are scored only once in Table I because they are the same word, but twice in Table II because they represent two different lexical notions. <sup>(B)</sup> There are more matches that involve all 6 languages in Table II because of cases like lexical notion #22, where neither of the Russian words has counterparts in each of the other languages, but the notion of "cooking" does:

English	German	Hindi	Marathi	Finnish	Russian
bake	backen	pakānā		paistaa	pec'
		bhājnā	bhājṇē		varit'

What these points illustrate is the fact that neither method immediately stands out as more accurate, or more useful.

The percentage of matches between individual languages and total resemblances in the data also varies depending on the method of scoring. Compare the following two Tables:

Table III: Percent of resemblances, out of 81 total (based on the data in Table I)

Russian to English	Russian to German	Russian to Hindi	Russian to Marathi	Russian to Finnish
60.5%	52%	43%	44.5%	37%

Resemblances between all 6 languages: 17%
Resemblances between all, and only, IE. languages: 8.5%
Resemblances between all languages except for one IE. Language: 2.5%

Table IV: Percent of resemblances, out of 57 total (based on the data in Table II)

Russian to English	Russian to German	Russian to Hindi	Russian to Marathi	Russian to Finnish
75.5%	66.5%	51%	54.5%	49%

Resemblances between all 6 languages: 30%
Resemblances between all, and only, IE. languages: 12%
Resemblances between all languages except for one IE. Language: 0%

Although the values in the Tables are different, their relationships remain fairly stable: Values in Table III are consistently lower than those in Table IV, but in both Tables matches between Russian and English are higher than others, and matches between Russian and Finnish are the lowest. Differences that stand out between the two methods of scoring are: 1) The percent of matches between all 6 languages, which in Table III is calculated at 17% and in Table IV at 30%; and 2)

The Percent of resemblances between all, and only, IE. languages which rises from 8.5% in Table III to 12% in Table IV. The significance of these differences, if any, has yet to be determined.

Regardless of the method used for scoring (i.e. counting hits for each lexical item or for entire notions), and in view of the fact that the relationship between the results of both methods remains stable, Greenberg's promise to predict which languages are more closely related than others can now be addressed. Do the numbers above clearly indicate that Russian is genetically closer to the other IE. languages than it is to Finnish? Indeed, the number of resemblances between Russian and Finnish are surpassed by the number of resemblances between Russian and any other IE. language. But not by much, and in fact as Table IV shows, the difference between Russian/Finnish and Russian/Hindi is only 2%. To add insult to injury, more matches between all 6 languages are found than between, and only between, the 5 IE. languages. These results are easily explained as borrowings between languages in close proximity to each other. But, such information is not available at this stage of the analysis, and would never be available if Mass Comparison were carried out on reconstructed languages whose history is not known.

There is one significant result, common to both methods described above: There are very few cases where all the languages, including Finnish but excluding just one IE. language, are found to correspond. Can this be interpreted as reflecting some kinship between the IE. languages? If all 6 languages were equally related then the number of matches involving any 5 languages should be similar regardless of which 5 are chosen. This is not so: Hits involving all 5 IE. languages surpass the number of hits that involve any other combination of 5 languages. Taking Mass Comparison a step further, by adding more languages, may shed more light on whether this figure is significant, or if it merely reflects a wishful and creative interpretation of the numbers involved.

### 3.2.2. STAGE II: WEEDING OUT “FALSE FRIENDS”

Etymology of Russian is instrumental in weeding out 12 lexical items that are found to have originated from sources other than inheritance. In some cases this process eliminates differences between Table I and II above. For example, in lexical item #16 both *plam'a* and *ogon'* are judged as resembling a synonymous word from another language. However, *plam'a* is shown to have resulted from borrowing while *ogon'* is found to be an inherited word. Similarly, *frukt* and *plod*, as well as *muskul'nyj* and *mys'ec'nyj*: In each case one synonym is found to have been borrowed and the other inherited. If Mass Comparison is to be used with languages whose histories are not already known, an awareness of this tendency may be helpful.

Etymological information of the remaining languages<sup>3</sup> at this stage of the analysis helped eliminate other resemblances that resulted from non-genetic origins. The remaining list (Appendix III) consists of the following: 1) Words that could not be eliminated because their origin could not be determined at this time due to lack of appropriate references (marked by an asterisk); 2) Words that could not be eliminated because their origin has not been determined in literature (also marked by an asterisk); and 3) Words that could not be eliminated because they are inherited. The matching words in Appendix III are therefore either genetically related, or cannot at this time be shown to be otherwise. The figures in Table V present the results using both methods described above: Counting each word as a separate hit on the one hand, and counting entire lexical notions on the other.

---

<sup>3</sup> Etymological data for German, Hindi, and Marathi, as well as some Finnish was kindly provided by Prof. Hock.

Table V: resemblances that were not eliminated from the list

Total Russian to any language		Russian to English		Russian to German		Russian to Hindi		Russian to Marathi		Russian to Finnish	
Words (W)	Notions (N)	W	N	W	N	W	N	W	N	W	N
38	30	17	16	19	18	18	17	17	15	11	11

Resemblances between all 6 languages:	Words: 1	Notions: 1
Resemblances between all, and only, IE languages:	Words: 11	Notions: 11
Resemblances between all languages, except for one IE. Language:	Words: 0	Notions: 1

The elimination of a large number, probably most, of the “false friends”, i.e. words that displayed similarities that were not indicative of genetic relationship, has served to bring the results of the two scoring methods much closer. 53% of the similarities that counted each word and 47% of the similarities across entire lexical notions are shown to have resulted from chance, borrowing, onomatopoeia, or nursery language. If counting similarities across notions seems to have fared better, it isn't by much. What these figures do indicate is that only an average of one lexical notion per language has more than one inherited synonym that is useful for this type of study. In other words, scoring by notions alone and ignoring multiple resemblances within them, would serve to overlook an average of only one significant item.

Echoing Peter's results, none of the resemblances between all 6 languages are proven to have resulted from genetic relationship. Only one such item (lexical notion #80, that of ‘wind’) remains in the final list of resemblances, but the origin

of the Finnish term *veto* has not yet been determined. Even if *veto* is found to have been inherited, only 1 out of 14 or 17 original resemblances, depending on the scoring methods described above, between all 6 languages survives.

As far as resemblances between the 5 IE. languages are concerned, about half of them turn out to be “false friends”: Table I shows a total of 21 matches between the 5 IE. languages (calculated as resemblances between all 6 languages, plus resemblances between all, and only, IE. languages); Table II shows this figure to be 24. After the process of elimination, in Table V, the remaining matches are counted as 12 (same for both methods of scoring). 43% of the lexical items, and 50% of the lexical notions scored as similar between all 6 languages are eliminated as not indicative of genetic relationship between these 5 languages, known to be related.

#### 4. CONCLUSION

An additional language, Russian, is added to the work of Peter and Hock in order to test Greenberg’s claim that Mass Comparison becomes more accurate as the number of languages is increased. Results from a total amount of 6 (English, German, Hindi, Marathi, Finnish, and Russian) languages are now measured up against Peter’s results using 3 (English, Hindi, and Finnish). Peter found 66% of total similarities to have been “false friends”. The present results are in fact slightly better: 53% of the similarities between separate words, and 47% of the similarities between whole lexical notions are determined to have been “false friends”. Alas, these figures cannot yet be celebrated as providing support for the method of Mass Comparison. Many of the Finnish words, those most likely to not be genetically related to the rest of the data, have not yet been fully investigated. It

would be premature to suggest that the accuracy of Mass Comparison has improved before the origin of more of the Finnish data can be determined<sup>4</sup>.

In general, Mass Comparison has not proven itself, at least not yet, as a reliable method for the classification of languages at a great time depth, nor for reconstruction. The methodology is difficult to replicate as it is lacking in clear criteria and often relies on fuzzy distinctions. Many inaccuracies are uncovered when etymology is consulted; inaccuracies that would not be exposed if there were no data available on the origin of the languages involved, as would be the case if Mass Comparison is used the way it was intended—with languages whose history is not known.

Intuitively, Mass Comparison seems quite attractive. The desire to look farther back into the history of languages than has been possible with more traditional methods is ever-present. Greenberg's method offers a strategy to circumvent the deficiency in lexical data from each individual language by the addition of lexical data from other languages. However, language is an unpredictable creature; what is known about a specific language and its history may not reflect what changes took place farther back. Therefore, if only limited data from each language is available, how much similarity to another language can be considered sufficient before their relationship can be determined? Mass Comparison attempts to answer this question, but the answer cannot (yet?) be confirmed.

---

<sup>4</sup> An attempt will be made to complete this stage of the study and present more complete results in the near future.



# PHYLUMPHILE OR PHYLUMFOE ?

Appendix I: List of all English and Russian terms.

#	ENGLISH	RUSSIAN
1	angry, irate, mad, raving, furious, enraged, vexed	serdityj, vspyl'c'ivnyj, gnevnyj, zloj, vozmusc'onnyj, yarostnyj, razdraz'onnyj, zlo, yarost', gnev, vozmusc'enije
2	arrive, come, reach	prijexat', prijti, pribyt', idti, dostignut', dobrat's'a
21	come	prijti
3	arrow	strela
4	ashes, cinders (cf. 28 "dust, powder")	pepel, zola
5	aunt	t'ot'a
6	bathe, wash	myt', kupat', stirat'
7	bear, carry, suffer, endure, cart	nosit', stradat', muc'atsa, terpet', povozka, kol'aska, telega, ruka (hand)
8	behind, in back	zad, popa, szadi, pozadi
9	blow (v.), puff, huff (cf. 80)	duf', dunovenije
80	wind (n.), air (cf. 9)	veter, vozdux
10	boil (v.), seethe (cf. 22)	kip'atit', varit', tus'yt'
22	cook (v.), bake, fry (cf. 10)	varit', pec', z'arit'
11	break (v.), burst, tear, split	razbit', vzorvat', rvat', razdelit'
12	breast, bosom, bust, boob, udder (cf. 17), tits	grud', b'ust, grudi, tit'ki, vym'a
17	chest (body) (cf. 12)	grud'
13	bring, take (to), fetch	nesti, vz'at', zabrat', brat'
14	broad, wide (cf. 32)	s'irokij
32	far, distant (cf. 14)	daleko, ot dal'onnyj (removed)
15	brother	brat
16	burn, blaze, fire, flame (cf. 48)	z'ec', goret', pylat', plam'a, ogon', kost'or (campfire)
48	light (n.), lamp (cf. 16, 67)	svet, fonar', ogon', lampa
67	shine (cf. 48)	svetit', blestit', cijat', jarkij, sijanie
18	chew	z'evat'
19	clean (v.), cleanse	ubirat', oc'isc'at', c'istit'
20	close (v.), lock, finish, end, shut (cf. 23)	zakryvat', zapirat', konc'at', zakanc'ivat', konec
23	cover (v.), deck (n) (cf. 20)	pokryvat', zakryvat', kryt'
24	cry (v.), scream, whine, shout, call; whimper	plakat', kric'at', xnykat', nyt', skulit', zvat'
25	dark, black, dusk, dismal	t'omnyj, temnota, t'ma, c'omnyj, mrac'nyj, tosklivyj, pasmurnyj
26	dirty, filthy, soiled	gr'aznyj, sal'nyj, skabr'oznyj
27	dog, hound, cur	sobaka, gonc'aja, p'os, ubl'udok
28	dust, powder	pyl', prax, poros'ok
29	ear	uxo
30	eye (cf. 66)	glaz
66	seek, look, search, see (cf. 30), view	smotret', videt', iskat'
31	fall, drop, plunge, tumble	padat', upast', ron'at', valit's'a
34	fear, fright, anxiety	strax, isprug, bojat's'a, bespokojstvo, ozaboc'ennost'
35	flea	bloxa

36	fruit	frukt, plod
37	girl, gal, daughter, maid(en)	devoc'ka, devus'ka, doc', doc'ka, deva, devica, sluz'anka
38	grass, hay	trava, seno
39	hair, mane	volos, griva
40	hit, beat, strike (cf. 44)	udar'a't', bit'
44	kill, murder, slay (cf. 40)	ubivat', ubijstvo, pogibnut', mertvit', m'ortvyj
41	horn	rog
42	house, home, hut, abode	dom, xiz'ina, z'ilis'c'e
43	husband, spouse (cf. 60)	muz', suprug
60	person, man, human being (cf. 43)	c'elovek, muz's'c'ina, muz'ik
45	know, understand	znat', ponimat', pon'at', podrazumevat'
46	laugh	smejat's'a, smex
47	leg, foot	noga, lapa
49	lip	guba
50	liver	pec'en'
51	mouse	mys'
51a	bat	letuc'aja mys'
52	narrow, tight (cf. 68, 71, 76)	uzkij, tugoij, tesnyj
68	short, brief, stubby (cf. 52, 71, 76, 84)	korotkij, kratkij, tolstyj
71	small, little, puny, (s)light, tiny, fine (cf. 71, 76, 84)	malenkij, nebol's'oj, melkij, s'c'uplyj, nic'toz'nyj, l'ogkij, kroxotnyj, kros'ec'nyj, mal'usen'kij
76	thin, meager, emaciated (cf. 52, 68, 71)	tonkij, skudnyj, xudoj, tos'ij, istos'onnyj, izmoz'd'onnyj, isxudalyj
84	young, new, fresh (cf. 68, 71)	molodoj, novyj, svez'yj
54	near, close (cf. 68)	blizkij, blizko
53	navel, bellybutton	popok, pup
55	neck (cf. 77)	s'eja
77	throat, gullet (cf. 74, 55)	gorlo, glotka
74	swallow (v.), gulp, gorge (cf. 77)	glotat', proglatyvat', xlebat'
56	nose, snout	nos, rylo, morda
57	old, ancient	staryj, drevnij, starinnyj
58	pain, suffering, soreness	bol', stradanije
59	penis, dick, prick, cock	xuj, xer, pipis'ka, xren
61	rain	doz'd'
62	red, ruddy	krasnyj, ryz'yj (red haired), rum'anyj
63.	root, wort	koren'
64	round, circular, rotund	kruglyj, puxlyj
65	saliva, spit, spittle, sputum	sl'un'a, plevat', plevok
69	shoulder	plec'o
70	sing, chant	pet'
72	sour, acid	kislota, kislyj
73	strong, muscular, tough, powerful, mighty	sil'nyj, krepkij, proc'nyj, mysec'nyj, muskul'nyj, mos'nyj, moguc'ij
75	testicle, ball	jaic'ko
78	two	dva
79	vagina, cunt, tail, pussy	pizda, vlagalis'e, manda
81	wing, pinion	krylo, s'estern'a

PHYLUMPHILE OR PHYLUMFOE ?

82	wish (v.), desire, hope, will	xotet', z'elat', nadez'da, vol'a
83	yesterday	vc'era

Appendix II: List of all Russian words that resembled any words from the other languages.

#	ENGLISH	GERMAN	HINDI	MARATHI	FINNISH	RUSSIAN
1	irate					serdityj
	furious		sarōṣ	ruṣṭa		yarost'
2	reach	erreichen				prijexat'
					yltää	prijti
4				phupētā		pepel
5	aunt	tante	kāki, čāci	kāki	tāti	t'ot'a
6					kylpeä	kupat'
7					kuljettaa	kol'aska
	cart	karren	kar	kar	koura	ruka
80	wind	wind	vāt	vāt	veto	veter
			vāyu	vāyu		vozdux
22	bake	backen	pakānā		paistaa	pec'
			bhājnā	bhājnē		varit'
12	bust	büste				b'ust
	tits			čūči		tit'ki
17		brust				grud'
14			čaurā			s'irokij
32	far	entfernt	dūr	dūr	etäällä	daleko
15	brother	bruder	bhāi, birādar	bhāū		brat
16	flame	flamme	jalānā	pōlṇē, jalṇē	palaa, loimo	plam'a
	fire		jvālā	jāl	palo	pylat'
			āg	āg		agon'
48	lamp	lampe			lamppu, lämpö	lampa
			jyōti	ujēḍ, jyōti		z'ec' [burn]
67	shine	scheinen				sijanie
18	chew	kauen	čabānā	cāvnē		z'evat'
19					siistiä	c'istit'
20	close					kryt'
24	cry	schreien	kikiyānā	kēkṇē	kirkua	kric'at', krik
			čik(h)/xnā			xnykat', nyt'
25	dusk	düster				tosklivyj
		schwarz				c'ornyj
			surmaī			pasmurnyj
				tāmas	tumma	t'ma
26	soiled					sal'nyj
27	hound	hund				gonc'aja
28		pulver			pöly, pulveri	pyl'

## HISTORICAL LINGUISTICS AND LEXICOSTATISTICS

29	ear	ohr				uxo
66	seek	suchen, sehen				iskat'
	view				tavoittel %la	videt'
31				padnē	pudota	padat'
34	fright	schrecken, furcht			aleta	valit's'a strax
			bhay	bhay, bāj		bojat's'a
35	flea	floh				bloxa
36	fruit	frucht				frukt, plod
37	daughter	tochter	dhiyā	dhida	tytār	doc'
				dhūv		deva
38					heinā	seno
39					harja	griva
40	beat					bit'
44	murder	morden	mār qālnā	mārṇē	murhata	mertvit'
41	horn	horn				rog
42	house	hous	ghar	ghar		xiz'yna
	home	heim	dhām	dhām	huone	dom
43					mies	muz'
60					mies	muz's'c'ina muz'ik
45	know	kennen	jānnā	jānnē		znat'
51	mouse	maus	mūs, mūsikā	mūṣak		mys'
51a		fleder% maus				letuc'aja m ys'
52	tight		taṅg		tarka	tugoij
71	light	leicht	laghu	laghu, lahān	lyhyt	l'ogkij
76	thin	dünn	tanu	tanu	hieno	tonkij
					ohut	xudoij
84	new	neu	nav, navā	navā	nuori	novyj
53				bōmbī		pup
77	gullet	gurgel	galā, kaṇṭh	ghātī	kurkku	glotka
74			harapnā			xlebat'
56	nose	nase	nāsā	nāsā	nokka	nos
62	ruddy					ryz'yj
64	circular					kruglyj
65	saliva		salil	salil		plevat'
73	muscular	muskulös				mys'ec'nyj muskul'nyj
	mighty	mächtig				moguc'ij
		mächtig				mos'nyj
78	two	zwei	dō	dōn		dva

# PHYLUMPHILE OR PHYLUMFOE ?

82	will	wollen		vañcchñē		vol'a, z'elat'
83	yester% day	gestern	kal, hāl	gēlā, kāl		vc'era

## Appendix III: List of Matches after Elimination of Borrowings, Nursery Language, and Onomatopoeia.

#	ENGLISH	GERMAN	HINDI	MARATHI	FINNISH	RUSSIAN
2					*yltää	prijti
4				*phupētā		pepel
80	wind	wind	vāt	vāt	*veto	veter
22			pakānā		*paistaa	pec'
14			*čaurā			s'irokij
15	brother	bruder	bhāi	bhāu		brat
16					*palo	pylat'
			*jvālā	*jāl		z'ar
			āg	āg		agon'
67	shine	scheinen				sijanie
18	chew	kauen	čabānā	cāvñē		z'evat'
19					*siistiä	c'istit'
25		*düster				tosklivyj
			*surmaĩ			pasmurnyj
				tāmas	*tumma	t'ma
28					*pöly	pyl'
29	ear	ohr				uxo
31					*pudota	padat'
					*aleta	valit's'a
34		schrecken				strax
			bhay	bhay, *bāj		bojat's'a
37	daughter	tochter	dhiyā	dhīda		doc'
				*dhūv		deva
44	murder	morden	mār dālnā	mārñē		mertvit', m'ortvyj
42	house	hous				xiz'yna
43					*mies	muz'
60					*mies	muz's'c'ina
45	know	kennen	jānnā	jāññē		znat'
51	mouse	maus	mūs, mūsikā	mūṣak		mys'
52					*tarka	tugoj
71	light	leicht	laghu	laghu, lahān		l'ogkij
76	thin	dünn	tanu	tanu		tonkij
84	new	neu	nav, navā	navā		novyj
77			*galā			glotka
56	nose	nase				nos
73	mighty	mächtig				moc'
78	two	zwei	dō	dōn		dva
82	will	wollen				vol'a

## LITERATURE

- Buck, Carl Darling. 1949. *A Dictionary of Selected Synonyms in the Principal Indo-European Languages*.  
Chicago: The University of Chicago Press.
- Collinder, Björn. 1977. *Fenno-Ugric vocabulary*.  
2nd ed. Hamburg: Buske.
- Décsy, Gyula. 1990. *The Uralic protolanguage: A comprehensive reconstruction*.  
Bloomington, In: *Eurologia*.
- Dolgopolsky, Aaron B. 1964. *A Probabilistic Hypothesis Concerning the Oldest Relationships among the Language Families in Northern Eurasia*.  
In: ed. V. evorokin T. L. Markey. *Typology, Relationship, and Time*.  
Ann Arbor: Karoma. 1986, pp. 34-35.
- Greenberg, Joseph H. 1957. *Essays in Linguistics*.  
Chicago: The University of Chicago Press.
- Greenberg, Joseph H. 1963. *The Languages of Africa*.  
Bloomington: Indiana University; The Hague: Mouton.
- Greenberg, Joseph H. 1987. *Language in the Americas*.  
Stanford: Stanford University Press.
- Hakulinen, Lauri. 1960. *Handbuch der finnischen Sprache*, vol. 2.  
Wiesbaden: Harrassowitz.
- Hock, H. H. 1991. *Principles of Historical Linguistics*.  
2nd ed. Berlin: Mouton de Gruyter.
- Hock, H. H. & Joseph, Brian D. 1996. *Language History, Language Change, and Language Relationship. An Introduction to Historical and Comparative Linguistics*.  
Berlin: Mouton de Gruyter.
- Itkonen-Joki. *Suomen kielen etymologinen sanakirja*.  
4 vols. (A-Teili). Vol. 1 ed. by Y. H. Toivonen, Vol. 2 by Y. H. Toivonen, Erkki Itkonen, and Aulis J. Joki, Vols. 3 and 4 by Itkonen and Joki. (Lexica Societatis Fenno-Ugricae 12:1-4.) Helsinki: Suomalais-ugrilainen seura, 1955-1969.
- Katzner, Kenneth. 1984. *English-Russian, Russian-English Dictionary*.  
John Wiley & Sons.
- Klein, Ernest. 1966. *A comprehensive Etymological Dictionary of the English Language*.  
Amsterdam: Elsevier Publishing Company.
- Newman, Paul. 1955. *On Being Right: Greenberg's African Linguistic Classification And The Methodological Principles Which Underlie It*.  
Bloomington: Indiana University.
- Oxford Dictionary On-line.
- Peter, Steven J. 1991. *Barking up the Wrong Family Tree? Greenberg's Method of Mass Comparison and the Genetic Classification of Languages*.  
B.A. Thesis. Department of Linguistics. UIUC.
- Rédei, Károly. 1986-88. *Uralisches etymologisches Wörterbuch*  
vols. 1-5 with continuous pagination. Vols. 1-3: Wiesbaden: Harrassowitz. Vol. 4-5:  
Budapest: Akadémiai Kiadó.
- Ringe, Donald R.. 1992. *On Calculating the factor of chance in language comparison*.  
In *Transactions of the American Philosophical Society*, 82:1.  
Philadelphia: American Philosophical Society.
- Ringe, Donald R. 1995. "Nostratic" and the factor of Chance.  
In *Diachronica* 12:1: 55-74.
- Ševoroškin, Vitaly. 1990. *The Mother Tongue*.  
In *The Sciences*. May/June: 20-27.
- Ševoroškin, Vitaly. and Markey T. L. Ed. 1986. *Typology, Relationship, and Time*.  
Ann Arbor: Karoma.

## PHYLUMPHILE OR PHYLUMFOE ?

- Starostin, S. A. *Comparative-Historical Linguistics And Lexicostatistics*.  
This Volume.
- Turner, R.L. 1962-1969. *A Comparative Dictionary of the Indo-Aryan languages*.  
London: Oxford University Press.
- Vasmer, Max. 1964. *Etymologi českij Slovar' Russkogo Jazyka*.  
Moskva: Progress.
- Wedgwood, Hensleigh. 1872. *A Dictionary of English Etymology*.  
New York: Macmillan & Co.





WHAT ARE SUFFICIENT CRITERIA FOR ESTABLISHING  
GENETIC RELATIONSHIPS AMONG LANGUAGES ?  
TESTING 'MASS COMPARISON'

Rajka Smiljanic

In this paper, I test Greenberg's methodology for establishing the relations between languages called 'mass comparison.' In order to do that, I provide the background concerning the comparative method and some details about Greenberg's approach. Then I outline Peter's ( 1991 ) study, and finally, I discuss the problems and the results obtained after applying 'mass comparison' to six languages including Serbo-Croatian.

1. THE COMPARATIVE METHOD

The comparative method has been used over the years to establish relationships between languages and to reconstruct the protoforms. During this time, a rigorous methodology has been established. This methodology consists of several steps. First, the assessment of languages, families, or phyla that need to be compared (based on the examination of basic lexemes and grammatical morphemes). Second, identifying of phonetic correspondences and diachronic sound changes along with the reconstruction of proto forms. Third, the interpretation of these forms which can lead to the reconstruction of the proto-languages and, finally, the etymological analysis of the lexemes and morphs.

However, before the reconstructed system is interpreted as a proto-language, other possible interpretations such as chance similarities, borrowings, nursery language, etc., have to be ruled out. These can be more easily recognized in cases where we have attested forms and written documents. Significant problems arise in

assessing the forms as we go further back in time and have no attested forms. Thus, the practitioners of comparative method emphasize, the more distant the linguistic relationships and the further back in time, the more rigorous procedures for substantiating these relationships are needed. Such rigorous methodology provides a tool for distinguishing between genetically related languages and those that are similar due to chance similarities or borrowings.

## 2. GREENBERG'S METHOD

Most often the formal apparatus of the Comparative Method has been applied to comparing a vast number of lexical items in two languages. However, Greenberg (1953) claims that in order to make language comparison more testable and provable, the comparison of several (many) languages is more fortuitous. He reasons that the probabilistic chance of coincidences or chance similarities is greater when comparing many morphemes in two languages than when comparing one morpheme in many languages. For this purpose, he uses a list of basic vocabulary which is less likely to be changed or borrowed and compares it in a great number of languages<sup>1</sup>.

Greenberg employed 'mass comparison' to classify languages of Africa and the Americas. The problem with these languages is that they are largely unwritten and they occur over a wide geographical area. In addition, Greenberg goes beyond individual present-day languages and families. He wants to establish larger families or phyla. In order to do that, Greenberg compares lexical items in many languages looking for the evidence of apparent phonetic similarities without establishing detailed correspondences. The similarities then suggest genetic closeness and possible affiliations of the languages being compared. In this way, he bypasses the establishment of regular correspondences as necessary for recovering relationships among languages--the ones that are more closely related and those that are more

---

<sup>1</sup>The question as to what constitutes basic vocabulary is still subject to debate.

distant--and furthermore, extends the pair-wise comparison to 'mass' comparison. This method, he argues, allows him to reduce chance similarities to minimal proportions: The greater the number of languages the more reliable the results. As a consequence, the addition of more languages should increase the validity of the results.

## 2.1. CRITICISM OF THE METHOD--RINGE (1992, 1995) AND PETER (1991)

Greenberg's approach provoked criticism from many linguists. Ringe (1992) claims that the addition of more languages that are being compared does not decrease but, in fact, increases the probability of the occurrence of chance similarities. Peter's (1991) study compares three languages with known histories, namely English, Finnish and Hindi, in order to test Greenberg's approach. By looking at 84 glosses in these three languages, and by scoring the phonetic similarities among them, Peter examines the applicability of 'mass comparison' and Greenberg's claim that the more closely related languages can be detected by 'eyeballing' the data and that furthermore, the addition of more languages (in this case comparison of more than two) decreases chance similarities.

Peter found that all three languages exhibited over 2% of similarities which is a percentage indicative of genetic relationship suggested by Greenberg (1987). However, on closer inspection, after consulting etymological dictionaries, Peter discovered that none of the correspondences between English and Finnish, and Hindi and Finnish were due to genetic relationship. Rightly, Hindi and English were scored as being more closely related. However, even here the method exhibited some difficulties. Some of the scored similarities were the consequence of borrowing. Furthermore, not all of the true cognates were discovered by 'mass comparison.' As far as the relationship between all three languages is concerned, the method has failed to detect the right degrees of closeness and relationships. It suggested that 11.9% of items were indicative of close relationship among these three languages which implies close genetic affinity among all three of them.

## 3. AMASSING FOR THE 'MASS COMPARISON'

Although Greenberg is vague in telling us what qualifies as 'massive' enough for the application of his methodology, three languages might not be what he had in mind. In order to bring this test closer to 'mass' comparison, Hock (work in progress) adds data from two more languages--Marathi and German. I provide the Serbo-Croatian data and, thus, this study of six languages might be closer to Greenberg's intent. Furthermore, these data enable me to test whether adding more languages decreases the chance similarities.

## 3.1 STEP I AND II

After translating the 84 glosses into Serbo-Croatian (which in itself presented problems since there is no one-to-one correspondence between the terms), I proceeded with the 'inspection method'<sup>2</sup>. This involved 'eyeballing' the data in an attempt to detect phonetic similarities between forms with similar meaning. This step turned out to pose numerous problems. Since Greenberg does not exactly specify or quantify what constitutes enough phonetic similarity to score two forms as possibly related, the method seemed fairly impressionistic. For example, I scored *strijela* and *arrow* as exhibiting phonetic similarity although they correspond in only one consonant and both have a glide.

Furthermore, sometimes I scored a Serbo-Croatian word as similar to a word in another language. But because this word of another language was so similar to the word in a third language, I would score that word as well to be similar to the Serbo-Croatian word, although I might not have scored them in isolation. Thus, in Table 1., I originally scored SCr and English words to be phonetically similar but I extended that to the other three languages based on their similarity with the English word:

---

<sup>2</sup>The translated items and the scored sets are supplied in the appendices.

Serbo-Croatian	English	German	Hindi	Marathi	Finnish
žvakati	chew	kauen	čabānā	cāvnē	

Table 1.

Another problem that I encountered was when a single term in Serbo-Croatian corresponded to more than one term in other languages. For example SCr *kuhati* is used for both 'to cook' and 'to boil' as is a German term *kochen* which I scored to be a cognate with the SCr one but only once were they scored with the English word *cook*. This was especially present in various terms for spatial relations and dimensions such as *thin*, *light*, *narrow*, *tight* etc., all of which could be translated as SCr *tanak*. Sometimes I included SCr lexemes that would not be literal translations of the English glosses such as *usta* 'mouth' which I scored with Hindi and Marathi *ōsth* although the term in English was *lip* (indicating the possible semantic shifts in meaning that would not be caught by the method). All this proved to be a problem in scoring the similarities and quantifying the result.

Furthermore, the knowledge of SCr morphology and some language history might have made some of my decisions arbitrary. One of the Serbo-Croatian infinitival endings is *-ti*. Despite my knowledge of this, I scored the items in Table 2 as being related. In these forms the variation in the first stop could easily be explained by sound changes. The second alveolar stop in SCr, however, is a part of a different morpheme that I compared to the second stop/fricative in English and German root morphemes:

Serbo-Croatian	English	German	Hindi	Marathi	Finnish
prati	bathe	baden			

Table 2.

Similarly, I knew that SCr *peći* 'to bake' originally had a velar stop *k* in it, which is preserved in the third person plural form *peku* and this probably influenced my scoring of this form with *bake* and *backen*. In the same way, I scored *doći* 'to

come' with *dākhaI* although I know that *do-* is a prefix added to the imperfective verb *ičī* 'to go' while at the same time I am not sure whether that is the case with the Marathi form.

In addition, the knowledge of some of the regular sound changes/laws, i.e. Grimm's Law, could possibly preclude scoring SCr *kriknuti* and English *cry*. However, keeping in mind that Greenberg usually works with languages which are not as well known as IE languages and their histories, I decided to ignore the knowledge described above for the purposes of this test. This furthermore reinforced my impression that it is very hard to be systematic and consistent in applying the 'inspection method.' All this shows that there are problems regarding the 'inspection method,' systematicity of scoring, and the delimiting of the semantic fields in which one looks for possible cognates.

### 3.2. STEP III<sup>3</sup>

The next step in my research was to consult the etymological dictionaries and actually confirm or reject the scored relations. This step uncovered that some of the similarities were due not to genetic relations but to onomatopoeic words, nursery words, borrowings and such. Furthermore, some of the true genetic relations among some items were left uncovered.

As a consequence of the above, the word for *aunt* was scored across all 6 languages although it is a nursery word<sup>4</sup>:

Serbo-Croatian	English	German	Hindi	Marathi	Finnish
teta	aunt	Tante	kākī / čāčī	āt, cultī / kākī	tāti

Table 3.

<sup>3</sup>I am grateful to Dr. Hock for providing me with the needed German, Hindi, Marathi, and Finnish etymologies.

<sup>4</sup>Actually, the English/German words are not nursery words but are borrowed from a language in which it was a nursery word.

The following was scored as a cognate in five languages although it turned out to be a consequence of onomatopoeia:

Serbo-Croatian	English	German	Hindi	Marathi	Finnish
puhati/ duvati	blow, puff	blasen pusten	phūkhñā dhaūknā		puhaltaa puhista

Table 4.

The method failed in discovering the similarities due to the borrowings such as:

Serbo-Croatian	English	German	Hindi	Marathi	Finnish
lampa	lamp	Lampe			lamppu

Table 5.

Some SCr words turned out to be unrelated to the given words but related to some others that diverged in meaning so that they were not included in the glosses: *iskati* is not related to English *seek* but is related to *ask* which was not provided and I decided not to include it myself. Similarly, *strah* is not a cognate with *fright* but is with *stretch*. *rumen/ruddy/red*, *kružni/circular/ring* and some others have the same relation as just explained.

The method failed to discover that *daughter*, *tochter*, *dhīyā* and *dhūv* are all related to each other and SCr *kćer*. Also that *znati/know/kennen/jännä/jäppä* are all derived from PIE \**gnō-*.

### 3.3 RESULTS<sup>5</sup>

The following are the percentages of scored relations between SCr and other language(s):

2-way relations:

- with English 2/84 or 2.3% with neither term actually indicative of genetic

---

<sup>5</sup>These results have to be taken with a grain of salt since some of the etymologies need to be double checked.

relationship;

- with German 4/84 or 4.7% with only one entry denoting true cognates;
- with Hindi 1/84 or 1.2%--this term being a false cognate;
- with Marathi 3/84 or 3.6% none of which reflect genetic relationship;
- with Finnish 6/84 or 7.1 % with only one set having a possibly common source.

3-way relations:

- with English and German 17/84 or 20.2% with five terms being actual cognate sets and one having a relation with only one language out of two;
- with German and Hindi 1/84 or 1.2%--this term being a false cognate;
- with Hindi and Marathi 8/84 or 9.5% with three actual cognate sets;
- with Marathi and Finnish 1/84 or 1.2%--the term being a false cognate;
- with English and Finnish 1/84 or 1.2%--false cognate;
- with English and Marathi 1/84 or 1.2%--false cognate;
- with English and Hindi 1/84 or 1.2%--false cognate.

4-way relation:

- with English, Hindi and Marathi 1/84 or 1.2%--false cognate;
- with English, German and Hindi 3/84 or 2.3% with one item being an actual cognate set;
- with English, German and Finnish 3/84 or 3.6% with only one set having two terms related to the SCr words;
- with Hindi, Marathi and Finnish 1/84 or 1.2%--false cognate;
- with German, Hindi and Finnish 1/84 or 1.2%--false cognate.

5-way relation:

- with English, German, Hindi and Finnish 3/84 or 3.6% with one set of true cognates and one having only one language related to SCr;
- with German, Hindi, Marathi and Finnish 2/84 or 2.3% none indicative of



genetic relation;

- with English, German, Hindi and Marathi 8/84 or 9.5% with four real cognate sets and two where only two languages were related
- with English, German, Marathi and Finnish 2/84 or 2.3% with one set having possibly two related forms to SCr;
- with English, Hindi, Marathi and Finnish 1/84 or 1.2%--false cognate.

6-way relation:

- with English, German, Hindi, Marathi and Finnish 12/84 or 14,3% with only four sets having actual correspondences across all languages but Finnish (possibly one Finnish cognate).

### 3.4. INTERPRETATION OF RESULTS

When looking at two-way relations, it appears that SCr is most closely related to Finnish (7.1 %) based on scored phonetic similarities. The relation is close even to Greenberg's 8% from 1966 (percentage of similarities sufficient for suggesting the close genetic relationship between languages) and well above 2% from 1987. These results also suggest that SCr is not closely related to English or Hindi, both of which scored less than 2% correspondences.

The three-way relation shows a drastic change in the results. According to the findings, SCr is most closely related to English and German (20.2%) and then to Hindi and Marathi (9.5%). If indeed these numbers are indicative of the closeness of relations among languages it seems that SCr cannot be grouped as closely related with any other combination of languages including Finnish (all below 2%), which contradicts the results from the two-way comparison (but improves in correct predictions).

Four-way relations are all below the proposed 8%, although the relations SCr / English / German / Hindi and SCr / English / German / Finnish are above 2%. Five-way relation shows close affinity (9.5%) between SCr, English, German, Hindi and

Marathi and 3.6% between SCr, English, German, Hindi and Finnish.

Six-way relation with 14.3% of scored correspondences (well above even the initially required 8%) wrongly suggests close affinity among all six languages. This becomes clear when the etymologies are uncovered. 8 out of 12 or 66.6% of suggested cognate sets reflect pure chance similarities, nursery language, borrowings and/or onomatopoeia. That is only provided that there is a possibility of Finnish forms being related to the other languages in the other four sets (I have to allow for this since I have not been able to uncover all the Finnish etymologies, although this is very unlikely). This is higher than Peter's result of 40% suggesting the close relation among three languages--English, Hindi and Finnish.

#### 4. CONCLUSION

The findings of this study suggest that the increase in the number of languages compared does not necessarily increase the reliability of the method nor does it decrease the chance similarities. Furthermore, it shows that Greenberg's methodology allows for some impressionistic interpretation of the data (the 'inspection method'). Without the rigorous statistical check on the amount of similarities necessary for establishing a relationship among languages, the quantitative requirements remain somewhat vague. Finally, without establishing regular sound correspondences, it is hard to rule out similarities due to other reasons than genetic ones and to establish valid groupings which might result in proposing wrong groupings.

APPENDIX 1

English	Serbo-Croatian
1. angry, irate, mad, raving, furious, enraged, vexed	ljut, gnjevan, srdit, ogorčen, lud, mahnit, razjaren, ozlovoljen
2. arrive, come, reach (cf. 21)	stići, doći, doseći,
21. come (cf. 2)	doći
3. arrow	str(ij)ela, strelica
4. ashes, cinders (cf. 28)	pepeo, žar, žeravica
28. dust, powder (cf. 4)	prah, prašina, puder
5. aunt	tet(k)a, strina
6. bathe, wash	kupati, prati
7. bear, carry, suffer, endure, cart	(pod)nositi, izdržati, trpiti
8. behind, in back	straga, iza
9 blow (v.), puff, huff (cf. 80)	puhati/duvati, dahtati
80. wind (n.), air (cf. 9)	vjetar, dah, zrak/vazduh
10. boil (v.), seethe (cf. 22)	(s)kuhati, vreti, kipjeti, ključati
22. cook (v.), bake, fry (cf. 10)	kuhati, peći, pržiti
11. break (v.), burst, tear, split	lomiti, kršiti, razbiti, trgati, derati
12. breast, bosom, bust, boob , udder (cf. 17)	prsa, grudi, dojke, njedra, sise, cice, vime

17. chest (body) (cf. 12)	prsa, grudi
13. bring, take (to), fetch	donijeti, nositi, zgrabiti
14. broad, wide (cf. 32)	širok, prostran
32. far, distant (cf. 14)	dalek, udaljen
15. brother	brat
16. burn, blaze, fire, flame (cf. 48)	gor(j)eti, plamsati, plamen, vatra, oganj
48. light (n.), lamp (cf. 16, 67)	svjetlo, luč, lampa
67. shine (cf. 48)	sjati, svjetliti
18. chew	žvakati
19. clean (v.), cleanse	(o)čistiti
20. close (v.), lock, finish, end, shut (cf. 23)	zatvoriti, završiti, zaključati, skončati, zabraviti
23. cover (v.), deck (n.) (cf. 20)	pokriti, prevući, paluba
24. cry (v.), scream, whine, shout, call, whimper	plakati, vikati, kriknuti, cviljeti, vrištati, zvati, cmizdriti
25. dark, black, dismal	taman, mračan, crn, tmuran
26. dirty, filthy, soiled	prljav, zamazan, blatan
27. dog, hound, cur	pas/ker, kuja, džukela
29. ear	uho
30. eye (cf. 66)	oko

TESTING 'MASS COMPARISON'

66. seek, look, search, see (cf. 30)	tražiti, iskati, gledati, vidjeti
31. fall, drop, plunge, tumble	pasti/padati, ispustiti, uroniti, prebacivati se
34. fear, anxiety, fright	strah, bojazan, tjeskoba, strava
35. flea	buha, uš
36. fruit	voće, plod
37. girl, gal, daughter, maid(en)	djevojka, cura, (k)ćer(ka)
38. grass, hay	trava, pašnjak, sijeno
39. hair, mane	kosa, dlaka, griva
40. hit, beat, strike (cf. 44)	udariti, tući, pogoditi
44. kill, murder, slay (cf. 40)	ubiti, umoriti
41. horn	rog
42. house, hut, home, abode	kuća, koliba, baraka, dom, boravište
43. husband, spouse	muž, suprug
60. person, man, human being (cf. 43)	osoba, čovjek, muškarac
45. know, understand	znati, razumijeti
46. laugh	smijati se
47. leg, foot	noga, butina, stopalo
49. lip	usna, usta [mouth]
50. liver	jetra
51. mouse a. bat	miš a. šišmiš

52. narrow, tight (cf. 68, 71, 76)	uzak, čvrst, tijesan, tanak
68. short, brief, stubby (cf. 52, 71, 76, 84)	kratak, nizak, sažet, tupast
71. small, little, puny, (s)light, tiny, fine (cf. 71, 76, 84)	malen, sitan, slab, lagan, tanak
76. thin, meager, emaciated (cf. 52, 68, 71)	tanak, mršav, iscrpljen
84. young, new, fresh (cf. 68, 71)	mlad, nov, svjež
54. near, close (Cf. 68)	blizu
53. navel, bellybutton	pupak
55. neck (cf. 77)	vrat/šija
77. throat, gullet (cf. 74, 55)	grkljan, grlo, jednjak, ždrijelo
74. swallow (v.), gulp, gorge (cf. 77)	gutati, daviti se, proždirati
56. nose, snout	nos, gubica, njuška, rilo
57. old, ancient	star, drevan
58. pain, suffering, soreness	bol, patnja, trpljenje, bolnost
59. penis, click, prick, cock	penis, kurac, kara
61. rain	kiša
62. red, ruddy	crven, rumen
63. root, wort	korijen
64. round, circular, rotund	okrugao, obao, kružni, zaobljen

# TESTING 'MASS COMPARISON'

65. saliva, spit, spittle, sputum	slina, pljuvačka
69. shoulder	rame, pleća
70. sing, chant	pjevati
72. sour, acid	kiseo
73. strong, muscular, tough, powerful, mighty	jak, snažan, mišićav, moćan, silan, jak, golem
75. testicle, ball	mudo, jaja, testisi
78. two	dva
79. vagina, cunt, tail, pussy	vagina, rodnica, pička, pica, pizda
81. wing, pinion	krilo, pero
82. wish (v.), desire, hope, will	željeti, čeznuti, žudjeti, nadati se, htjeti, volja
83. yesterday	juče(r)

## APPENDIX 2

Serbo-Croatian	English	German	Hindi	Marathi	Finnish
1. ljut	wild	wild	lāl		
gnjev	anger				
razjaren	irate	rasend	nārāz	ruṣṭa	raivoisa
2. doći				dākhal	
21. poći			pahūcnā	pōhācṇē	
3. strijela	arrow		tīr, šar	tīr, šar	
4. žar	ashes	Asche	chār		
pepeo				phupētā	poro

28. puder	powder	Puder	dhurrā	pūḍ	pulveri
prah				raj	
5. teta	aunt	Tante	kākī/čācī	kākī	tāti
6. kupati					kylpeā
prati	bathe	baden			
7. trpiti		tragen			
8. straga, iza		zurück			
9. puhati / duvati	blow, puff	blasen, pusten	phūkhnā, dhaūknā		puhaltaa, puhista
80. vjetar, vazduh	wind	Wind	vāt, vāyu	vāyu	veto
10. kuhati		kochen	khaulnā	kaḍh(av)ṇē	kiehua, kuohua, kiehutta
kipiti					kypsyttā
22. kuhati	cook	kochen			
peći	bake	backen	pakānā		paistaa
pržiti	fry	braten			
11. derati, trgati	tear	brechen	tōrnā		rikkoa
12. prsa, grudi	breast, bosom	Brust, Busen			
cice, sisa				čūcī	
17. prsa, grudi		Brust			
32. dalek			ḍūr	ḍūr	



TESTING 'MASS COMPARISON'

15. brat	brother	Bruder	bhāī, birādar	bhāū	
16. plamen	blaze, flame	flammen, Flamme		pōlņē	polttaa, palaa
48. lampa	lamp	Lampe			lamppu
luč	light	Licht			lyhty
67. sjati	shine	scheinen			
18. žvakati	chew	kauen	čabānā	cavņē	
19. čistiti	clean				siistia
24. kriknuti	cry	schreien	čik(h)/xnā	kēkņē	kirkua
25. taman	dismal	dunkel		tāmas	tumma
cm		schwarz			
26. prljav	dirty	dreckig			
27. ker	cur	Köter	kuttā, kukkur	kukar, kutrā, kukkur	koira, rakki
29. uho	ear	Ohr			
30. oko	eye	Auge	ākḥ		
66. iskati	seek	suchen, sehen			
31. padati			paṛnā	paḍņē	pudota
34. bojazan			bhay	bhay, bāj	
strah	fright	schrecken			
36. plod			phal	phaḷ	

37. djevojka			dhīyā	dhūv	
38. trava	grass	Gras			
sijeno					heina
39. kosa	hair	Haar	kēš	kēs/kēs̃	hius
griva					karva
40. tući				ṭicṇē	
44. umoriti	murder	mordern	mār ḍālnā	mārpē	musertaa, murhata
44. kuća	house	Haus, Hutte	kōṭhā	kōṭhā	koti
dom	home	Heim	dhām	dhām	
43. muž					mies
60. muškarac					mies
45. znati	know	kennen	ǰānnā	ǰānnē	
46. smijati se			hasnā	hā/ aspe	
49. usta			ōṣṭh	ōṣṭh	
50. jetra			ǰīgar	yakrt	
51. miš	mouse	Maus	mūsikā, mūs	mūṣak	
52. tanak		eng	taṅg		tarka
68. kratak		kurz			
71. lagan	light	leicht	laghu	lahān, laghu	lyhyt
76. tanak	thin	dunn	tanu	tanu	

TESTING 'MASS COMPARISON'

84. nov	new	neu	nav, navā	navā	nuori
svjež	fresh	frisch			
53. pupak	belly- button			bōmbī	
77. grlo	gullet	Gurgel	galā, grasikā	galā	
74. gutati	gulp		gatakānā		
56. nos	nose	Nase	nāsā	nāsā	nokka
njuška		Schnauze	nāk	nāk	nokka
59. penis	penis	Penis			penis
62. rumen	ruddy	rot	arun	arun	
64. kružni, okrugli	circular	kreis- förmig			
65. slina	saliva		salil	salil	sylki
73. mušićav	muscular	muskulos			
jak		wacker			
75. testis	testicle				
78. dva	two	zwei	dō	dōn	
79. vagina	vagina	Vagina			
81. pero			par		
82. volja	will	wollen			
83. jučer	yesterday	gestern			

## LITERATURE

- Greenberg, Joseph H. 1957. *Essays in Linguistics*.  
Chicago University Press.
- Greenberg, Joseph H. 1963. *The Languages of Africa*.  
Bloomington: Indiana University ; The Hague: Mouton.
- Greenberg, Joseph H. 1987. *Languages in the Americas*.  
Stanford: Stanford University Press
- Hock, H. H. & Joseph, Brian D. *Language History, Language Change and Language relationship. An Introduction to Historical and Comparative Linguistics*.  
Berlin: Mouton de Gruyter.
- Peter, Steven J. 1991. *Barking up the Wrong Family Tree? Greenberg's Method of Mass Comparison and the Genetic Classification of Languages*  
B.A. Thesis, Department of Linguistics. UIUC.
- Ringe, Donald R. 1992. *On Calculating the factor of change in language comparison*.  
In Transactions of the American Philosophical Society, 82:1.  
Philadelphia: American Philosophical Society.
- Ringe, Donald R. 1995. "Nostratic" and the factor of Chance.  
In Diachronica 12:1. 55-74.

## ETYMOLOGICAL DICTIONARIES

- Buck, Carl D. 1965. *A Dictionary of Selected Synonyms in the Principal Indo-European Languages*.  
2nd ed. Chicago: The University of Chicago Press.
- Collinder, Bjorn. 1977. *Fenno-Ugric vocabulary*.  
2nd ed. Hamburg: Buske.
- Decsy, Gyula. 1990. *The Uralic protolanguage: A comprehensive reconstruction*.  
Bloomington, IN: Eurolingua.
- Gluhak, Alemenko. 1993. *Hrvatski Etimoloski Rjecnik*.  
Zagreb: August Cesarec.
- Hakulinen, Lauri. 1960. *Handbuch der finnischen Sprache, vol 2*  
Wiesbaden: Harrassowitz.

Itk.-Joki= *Suomen kielen etymologinen sanakirja*.

4 vols. (A-Teili). Vol. 1 ed. by Y. H. Toivonen, Vol. 2 by Y. H. Toivonen, Erkki Itkonen, and Aulis J. Joki, Vols. 3 and 4 by Itkonen and Joki. (Lexica Societatis Fenno-Ugricae 12:1-4)  
Helsinki: Suomalais-ugrilainen seura, 1955- 1969.

Klein, Ernest. 1966. *A Comprehensive Etymological Dictionary of the English Language*.

2 vols. Amsterdam: Elsevier Publishing Comp.

Rédei, Károly. 1986-88. *Uralisches Wörterbuch*, vols. 1-5 with continuous pagination.

Vols. 1-3: Wiesbaden: Harrassowitz. Vol. 4-5: Budapest: Akadémiai Kiadó.

Turner, R. L. 1962-1969. *A comparative dictionary of the Indo-Aryan languages*.

London: Oxford University Press.



**Part III:**  
**Calculating Language Relationship.**





# CALCULATING LANGUAGE RELATIONSHIPS AND PATHS OF DISPERSAL IN EURASIA DURING THE LAST 100,000 YEARS, USING THE LANGUAGE MODEL

Harald Sverdrup

## 1 Introduction

This study arose as a by-product of models for population dynamics and models for soil fertility, biodiversity and forest growth developed by the Biogeochemistry Division at the Department of Chemical Engineering, Lund University, Sweden. When doing mathematical modeling and systems analysis of environmental pollution impacts on ecological systems across Asia, we discovered that past climate changes and fossil pollen records could be used to test the predictive capability of mathematical ecological models for the effects of global climate change. Being used to make population models, it was only natural to experiment with modelling of human populations. Once we were able to use ecological factors for population control, the next natural step was to give them a language and study where the different languages would go.

## 2 Objectives and issues

The objective of this study is to mathematically model the geographical and temporal language distribution patterns in Eurasia as a result of major movements of paleolithic and mesolithic peoples in Eurasia during the time period from 100,000 BC to 10,000 BC, and on this overlay the effect of the neolithic transition 10,000 BC to 2,000 BC causing a demic inflation of the population sizes. The model is based on ecological mechanisms for driving the population transfer. The calculations are analyzed by using linguistic, genetic and archaeological information. The model will be used to reconstruct language dispersal, origins and relatedness. It will test the possibility of concepts like Nostratic, Austric, Dene-Sino-Caucasian. The feasibility of several issues can be

assessed with the LANGUAGE model such as, 1: if prehistoric language dispersal and geographical distribution can be reconstructed using mathematical models, 2: if the pre-colonial genetic pattern of the world's populations can be explained mainly by demic processes starting in Africa more than 100,000 years ago, 3: if the invention of agriculture in the Middle East and central China form plausible mechanisms for explaining historic pre-colonial language pattern of Eurasia and 4: if the language phylae Nostratic, Austric, and Dene-Sino-Caucasian are valid genetic language nodes that would be supported by the LANGUAGE model calculations.

### 3 Modelling philosophy and principles

All processes of change require per definition a driving force. This is also true for language change. Renfrew (1987, 1989), truly states that "without a true mechanism and a driving force, there can be no change", pertaining to linguistic or cultural change, this is a basic principle very well known in basic natural sciences and engineering, indeed, it is the Second Law of thermodynamics, and has universal validity. When a language has spread to an area or overlays the established language in a region, then this can only happen if there is a valid ecological or political mechanism for the transition. Renfrew (1990) has stated that the earlier applied equation "culture equals population equals language" should be abandoned in favor of a differential model stating that the change in one system will induce change in connected systems: "change in culture is proportional to change in demography is proportional to change in language". A process view implies a mechanism of change, depends on state variables for the system and is generally valid. The state of the system would be characterized by the state of the culture in terms of social order, demography and ecological adaption. The mechanisms that act are dependent on the properties of reacting medium and not dependent on location in time and space as such. Such a model has the advantage that it is generally valid and applicable when it can be properly parameterized and the coefficients estimated, but the drawback is that it is in differential form, and requires mathematical manipulation when used.

The principle applied in the model is illustrated in Figure 1. The basic equation for demic diffusion into an area such, is the general equation for diffusion and simultaneous reaction of the diffusing substance, the **equation of continuity**, derived from the principle of mass conservation in a system (Bird et al., 1960; Ammerman and Cavalli-Sforza, 1984). It describes the transfer of two substances through a volume element fixed in space, in our case the transfer of two types of population through a land surface element fixed geographically. Population growth in each geographical element is regulated by birth and death rates. Population in the geographical element is also changed when individuals enter the element from another element or when they leave it. Under conditions where a neolithic population is diffusing into the area, the diffusion of mesolithic hunter-gatherers out of the area will be proportional to the total number of people in the area. The diffusion of agriculturalists into the area will depend on the difference between the population density behind the wave front and that before it. Mostly the density of the hunter-gatherer population will be negligible as compared to the neolithic population, but there are some important exceptions. In cold areas will the fertility of the land decrease, and the saturation population density for neolithics will be lower. At some point it will sink to the level of the local mesolithic population, at which point the diffusion will come to a halt. There are also locations where fishing resources or local abundances of game allowed the density to become higher than normal. In such areas, the demic diffusion will be slower or halted. We assume that the farming population, utilize the landscape for hunting just the way the mesolithic population would do in addition to their farming activities. Thus within a cell, the mesolithic population will be subject to both adsorption into the farming population, and a large urge to leave for other hunting-grounds.

The LANGUAGE model was set up to calculate the initial colonization of the earth by modern man from 140,000 BP to the advent of agriculture, and the replacements caused by that event. The model has the following equations:

- Diffusion equation coupled with population growth:
  - Population growth kinetics based on Michaelis-Menten and linear mortality
  - Surface roughness adjustment of diffusion rates
  - Climate adjustment of all rates
- A vegetation net primary production model, based on soil properties and climatic conditions
- A fauna bulk production model, based on vegetation cover.

The primary invasion of biomass is assumed to follow clearance of land from ice by instant seeding in a zone adjacent to the nearest vegetation. The rate has been measured for post-glacial conditions at several locations. The ecological factors enter the calculations by the growth rate coefficients and the mortality coefficients. The factors are affected by temperature and annual rainfall, the transfer coefficient and diffusivity, are also affected by the roughness of the landscape. For the establishment of a functioning vegetation and fauna, the temperature dependence of growing trees and shrubs was used, and accordingly the temperature dependence of the lowest trophic level of the mesolithic ecological system. For the hunter-gatherer populations, the saturation density is dependent on the production of prey.

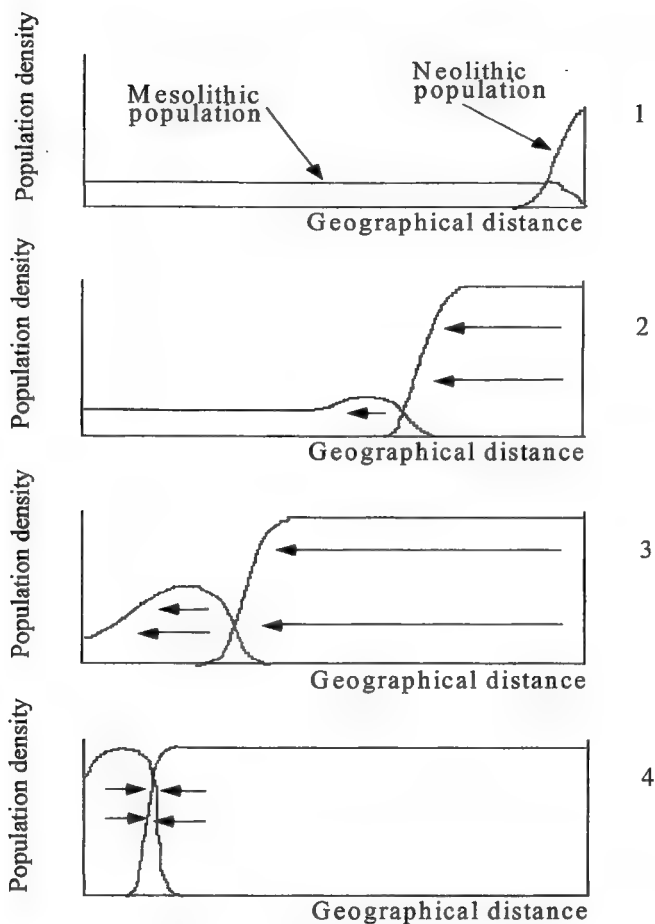


Figure 1. *The evolution of demic diffusion over time. The figure show how the wave of advance of the population density for farmers move in a front at constant speed. A part of the hunter-gatherer population is assimilated, the rest is displaced. Over time, the displaced population builds up a counterpressure to the advancing wave, if the displaced population can maintain the population density with modified food supply methods.*

Soil fertility is calculated as a function of soil moisture and chemical weathering rate, climate is expressed in terms of temperature, annual rainfall and annual precipitation surplus.

Language isogloss retention is calculated, assuming a glotto-kinetic model, using loss coefficients dependent on population density. The model is used to calculate the results in terms of:

- First date of arrival of wave of advance (1000 years BC)
- Geographical area covered as a function of time for
  - Huntergatherer population ( $\text{km}^2$ )
  - Neolithic population ( $\text{km}^2$ )
- Population density in each grid (persons  $\text{km}^{-2}$ )
- Total integrated population number for
  - Huntergatherer population (persons)
  - Neolithic population (persons)
- Fraction of original language remaining (%)

Within a short time the gradient at the wave front will reach a steady state, and the wave propagation velocity will depend only on the difference between the population density before the front and after it, however adapted regionally by landscape diffusivity and ambient temperature.

#### 4 Initial and boundary conditions

There are a number of dated events that can be used to constrain the model. The calculations rest partly upon soil data sampled throughout Far East Asia on a  $0.5^\circ$  by  $0.5^\circ$  grid, in Russia on a 150 km by 150 km grid, in Europe on a 50 km by 50 km grid and in Africa and America on a polygon basis approximately comparable to a  $2^\circ$  by  $2^\circ$  grid. The calculations are carried out in

two phases. Phase 1 comprise the initial peopling of Eurasia by modern man and spans the time period 100,000 BC to 8,000 BC. It is initialized with 10,000 individuals in Northeast Africa in Southern Ethiopia, on January 1., 140,000 BC, all having uniform language, new brains and full of initiative. Phase 2 comprise the neolithic transition in Eurasia during 8,000 BC to 1,500 BC. The demic diffusion due to neolithic transition is started in the Far East in the Chinese Central Yellow River Vally approximately 7,500 BC, at the same time in southern China in Fukien. It is also started in Jericho 8,800 BC, creating secondary initiation centres by 6,800 BC in different parts of the Fertile Crescent and Anatolia.

Event	Date
African migration to Near East	105,000-95,000 BC
Migration from Central Asia to Northern East Asia	50,000 BC
Migration from Central Asia to Southern East Asia	60,000 BC
Modern man replace neandertals in Europe	45,000-35,000 BC
Migration into Australia	40,000-30,000 BC
Amerindian migration to America	25,000-15,000 BC
Eskimo-Aleut migration to America	15,000-12,000 BC
Na-Dene migration to America	11,000-9,000 BC

Table 1. *Archaeological dates are used to check the performance of the model.*

#### 4.1 Observed dates of arrival

Archaeology and history form the ultimate data backdrop for checking the theories. Archeological artifacts do not carry any indication of the language of their makers, unless inscribed in written language, but still can give indications of when a certain type of cultural change took place. Approximately 12,000 BC, wheat was harvested systematically in the Natufian culture in southern Palestine. Wheat is cultivated on a larger systematical scale in the Levant and in the Northern part of the Fertile Crescent around 8,000 BC. Around 8,000-7,000 BC, initializing agricultural nuclei was established around Catal Hüyük in the Conya

plain in present Turkey (Proto-Indoeuropean), the Natufian culture centered on Jericho and Wadi Arabi (Proto-Afroasiatic), maybe at Cayonu in eastern Turkey (Proto-Caucasian, Hatti), around Ali Kosh in the Southern Zagros Mountains (Proto-Elamo-Dravidian) and maybe also at Jeitun in Transoxania (Proto-Ugric-Altaic). Agriculture based on millet and later wheat started as an independent invention in Northern China before 6,500 BC. At the same time, wet rice cultivation was initiated in the central Yang Tze River region of China. This was a development initiated from the Fukien-Tonkin area in Southern China around 8,000-7,500 BC, where horticulture was the initial form of agriculture. The introduction of agriculture into the Indian subcontinent probably occurred from Anatolia (Renfrew, 1989), entering India around 4,500 BC. The archaeological data was taken from Cavalli-Sforza (1988); Fagan, (1990); Laričev et al., (1988); Howells, (1988); Chang, (1983). Important corroborative data has come from different types of genetic research on the populations of the world. One approach uses present populations, and analyses a different number of genetic markers. The relatedness of different populations was quantified using a number of different genetic markers by Cavalli-Sforza, et al., (1988). The genetic tree is shown in Fig. 5. Whereas political mechanisms and elite dominance can change language, only quantitatively significant population migrations and demic expansions show up in the genetic information.

## 5 Results

### 5.1 Paleolithic Eurasia

The calculated situation in Eurasia during the paleolithic timeperiod from 100,000 BC to 20,000 BC is shown in Figure 2. The map show the position of the front of the wave of advance at different times. The lines are isochrons for the first appearance of modern man throughout Eurasia in thousand years BC. The black areas represent areas under ice. The glaciation situation has been simplified here in order to allow showing how modern man initially populated Eurasia in the period from 140,000 BC to 10,000 BC. The first divide between the population diffusing southeast of Himalaya and the other moving north and west creates the



divide between Austric-Indo-Pacific and Australian languages on one side and Nostratic, Amerind and Dene-Sino-Caucasian languages on the other. This divide occurred 60-70,000 years ago according to the model.

Nostratic and Dene-Sino-Caucasian divided as separate languages from their common Eurasian ancestor in central Asia 40-50,000 years ago. The period from 67,000 to 58,000 BC was among the coldest during the glacial period, and at this time Central Europe was probably a rather sterile polar desert. This changed after 45,000 BC when a short period of somewhat warmer weather followed. The northern part will eventually develop into a Dene-Sino-Caucasian group in the West and North and and Nostratic languages in the south and northeast, the southern movement continues south to become Austric, Indo-Pacific and Australian languages. The diffusion through Siberia came in the time period 50,000-30,000 BC. After 40,000 BC this region was very cold, and only a small number of people may have passed through in a second movement after 40,000 BC, subsisting on reindeers. The population density calculated for this area is small. The implication is that a restricted part of the genetic signal may have passed through, accounting for the east-west difference in genetic markers.

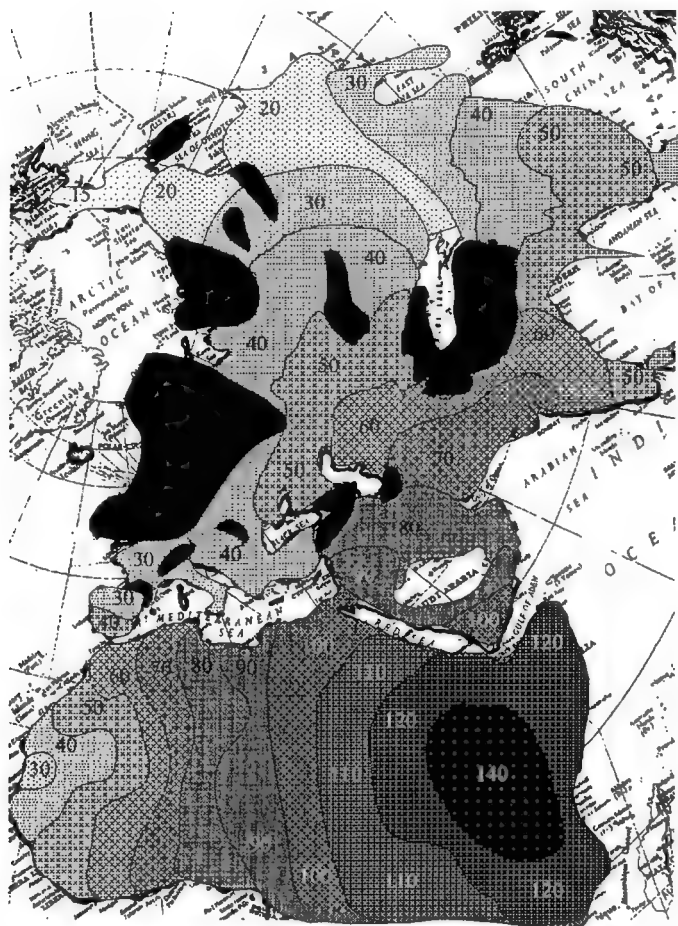


Figure 2. The calculated situation in the timeperiod 100,000-15,000 BP. The map show the calculated first appearance of modern man in Eurasia. Average icecover in 40,000-20,000 BP is shown, but remember this varied much during the period modelled. Numbers represent thousands of years BP. The contours of Eurasia have been distorted because of glaciation and depression of the sealevel.

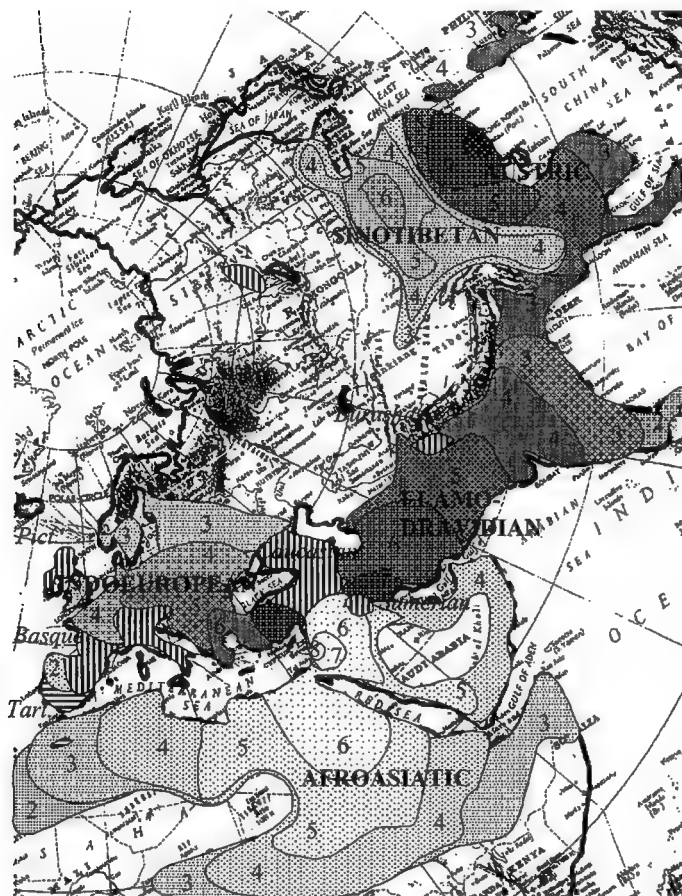


Figure 3. Calculated expansion due to demic diffusion driven by agriculture for Indo-European, Afro-Asiatic, Ugric-Altaic, Elamo-Dravidian Austric and Sino-Tibetan for the time period 8,000-2,000 BC.

Proto-Dene-Sino-Caucasian is originally a language of the western Eurasian reindeer hunters, which at the end of the glaciation after 30,000 BC spread eastward across Russia and Siberia. Initial settlement of Europe occurred approximately 50,000 BC from the East. In 32,000 BC, a period of very cold climate started which lasted till 23,000 BC, the Wurm-Weichsel glacial

maximum. This maximum transformed the Central European fringe to the icesheet to a very cold polar desert. The paleolithic resettlement of Europe after the retreat of ice approximately started maybe as early as 23,000 BC, but gained momentum 15,000-17,000 BC due to significant warming of the climate, when the polar desert disappeared. The upper paleolithic had been based on reindeer and its hunting grounds on the tundra relatively close to the continental icecap. The Younger Dryas, a cold period lasting from 8,800-8,300 BC, had a particularly profound effect on the conditions. It caused the already established northern boreal forest zone in Europe to revert to cold tundra and arctic desert. This reduced the population density to a very low level. During the Allerød oscillation 10,200-8,800 BC the climate improved over larger parts of Europe, and the reindeer economy pulled north with the ice and prey and survived for a period in the Allerød-Lyngby culture.

The eastern Sino-Caucasian group, what later became Sino-Tibetan, Yenesseian and Na Dene expanded eastward through Siberia, and also here acquired their morphological characteristics from exposure to extreme cold climate. During the last cold period of the early post-glacial 25,000-19,000 BC, these populations were shifted southward, thus establishing themselves in Northern China. This implies that the model suggests that the ancestors of the Sino-Tibetan speaking Chinese came to China from the very cold north approximately 12,000 BC. The cold period of the third Weichsel-Wurm glacial maximum 26,000-15,000 BC moved the tundra border south, and possibly pressed population groups in Europe and Central Asia farther south. Something similar but less severe occurred during the more recent Older Dryas and Younger Dryas cold periods. In Europe, this implied that the population density fell to very low level, and large regions being polar desert would have been uninhabitable. As the temperature rose after 21,000-15,000 BC, there would have been a reflux of Caucasian languages northward into Europe. There was a reflux from Iberia and Northern Africa along the Atlantic border towards the north, and another reflux started from Anatolia going northwest. According to this scenario, the western branches that originally split from Proto-Caucasian 30,000-40,000 years ago, divided internally into early

forms of Basque, Iberian and Pictish after 15,000 BC. The Eastern European branch of Caucasian represented by Proto-North Caucasian split from the Central Caucasian group containing Lemnian, Camunian, Rätian, Elymian and Etruscan approximately 12,000-15,000 BC according to our estimates.

## 5.2 Neolithic Europe and Africa

Initial language positions in Central Asia at the inception of agriculture is based partly on guesses and partly on modeling for the period 120,000 BC to 7,000 BC using the model. The calculated situation for the time period from 7,000 BC to 2,000 BC is shown in Fig. 3. Nostratic has completed the process of dividing into Afro-Asiatic, West-Nostratic comprising Indo-European and Kartvelian, and East Nostratic comprised of Elamo-Dravid, Ugric and Altaic. The transition to agriculture creates a wave of advance travelling out radially from the fertile crescent.

In 3,000 BC, Nostratic reaches the Atlantic Ocean and the Northern border of agriculture. Etruscan gets confined in the mountain valleys of the Central and North Appennine Mountains in Italy. Basque, Pict and Etruscan occupy confined enclaves. Iberian may also be such an enclave, but the model cannot distinguish between this alternative and Iberian being of North African derivation. The model calculations show that Caucasian groups in the fertile crescent are completely confined by Nostratic languages to the West, South and Southeast, In the north the steppe begins. Hence an expansion of Caucasian has no outlet to any significant amount of territory fit for agriculture. Proto-Elamo-Dravidian diffuses out from the inner eastern shore of the Persian Gulf together with agriculture starting at the same time as the other Nostratic languages. The languages that moved north and east will develop into Dravidian. It has been ascertained that Elamitic languages were once spoken over most of the Iranian plateau in early historic times, before 3,000 BC (Lamberg-Karlovsky, 1978). According to the LANGUAGE model, approximately 3,500 BC the wave of agriculturalists speaking Proto-Dravidian language meets a comparable wave of advance of peoples speaking Proto-Austroasiatic in the middle of India. During Chinese Han times, the first census

was held in China, the result was 52 million, suggesting that the population for the whole area was 65-75 million; the model predicts 62 million. The model estimates that the world had 400 million inhabitants in 200 BC, excluding America. The correlation between observed and calculated first date of arrival was  $r^2=0.91$  and the standard deviation  $\sigma = 5,100$  years, roughly equivalent to an accuracy of  $\pm 20\%$  on the calculated value. The observed correlation may reflect that the model does incorporate the most important driving factors.

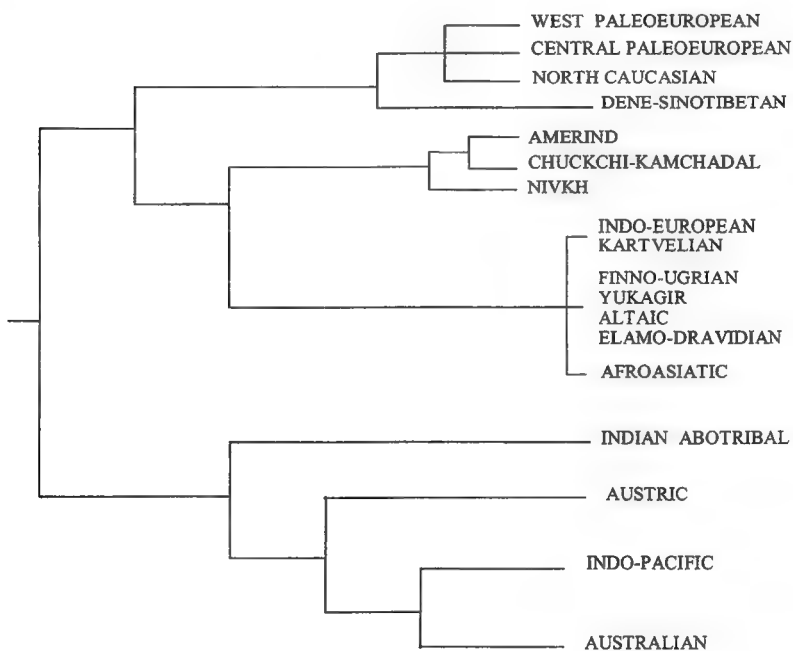


Figure 4. The calculated interrelation between the major language phylae of Asia, Europa and America using the LANGUAGE model.

Datings of first pottery, domesticated animals and a semi-neolithic type of culture in central Sahara overthrow the simple assumption that all Afroasiatic languages expanded from Jericho with agriculture (Kuper 1979). The archaeological data suggests an expansion of a pottery culture straight east and west starting 9,000 BC from the Hoggar-Tibesti region. The data suggest that the Mahgreb region was continuously populated with the peoples that were Caucasoid bearers of the Capsian culture, whereas south of the Atlas, the present desert, then steppe was inhabited by peoples of African stock (Nilo-Saharans). The Capsian cultural bearers are with all probability the descendants of the original Cro-Magnon population, the Basque are believed to be their descendants. It is probable that the initial split of Afroasiatic took place in Northeast Central Sahara, and expanded outward as Chadic, Cushitic, Proto-Berber and Semitic. It must then have been Old Semitic that expanded north-, east- and southward from Jericho, but the reflux back into North Africa moving over the semi-neolithic culture there brought the Coptic-Berber branch. By the time the expansion would have reached Libya by approximately 4,500 BC, the severe drought there would have effectively have limited the impact of the demic expansion to a small coastal band of North Africa. The early neolithic culture in the Sahara would have been cut off by the desert like an island. The neolithic culture reached the Canary islands by this process by 3,000 BC.

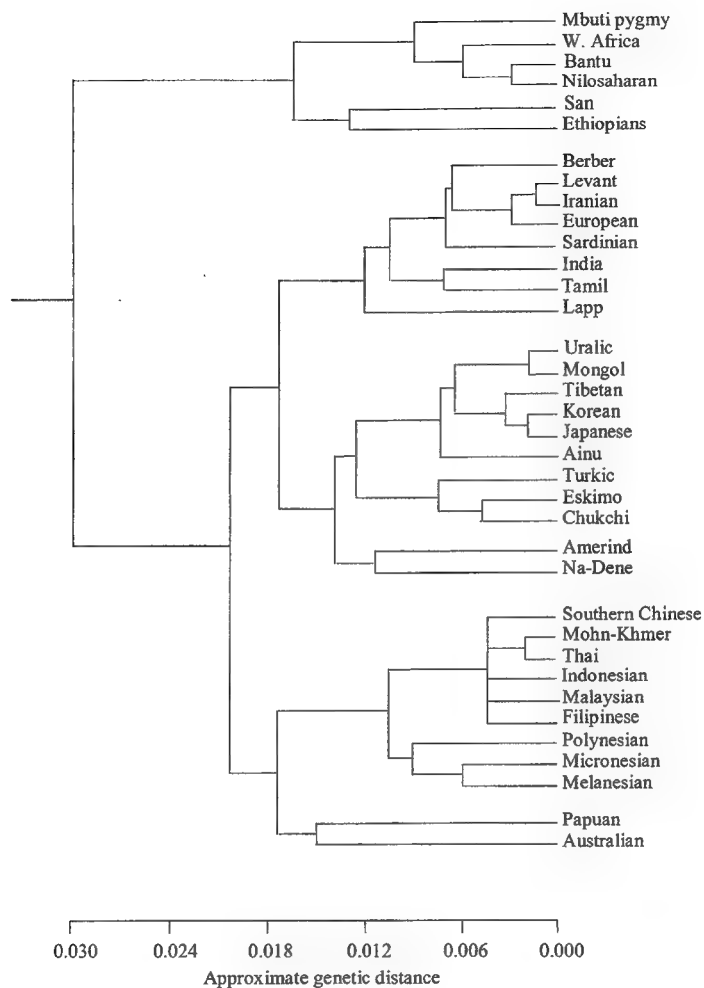


Figure 5. Genetic relation tree for populations of the world according to Cavalli-Sforza, Piazza, Menozzi and Mountain (1988) and Cavalli-Sforza (1991).



### 5.3 Neolithic East Asia

Fig. 3 shows the calculated demic diffusion initiated by the inception of agriculture. For the calculation two primary focuses were used, and a third secondary was allowed to form in the Central Yang Tze Valley. Austric has expanded with rice agriculture into Indo-China and Eastern India. The western wave of advance meets with the eastern wave in Eastern Central India. The neolithic transition occurs in northern China through the cultivation of foxtail millet. North Chinese meets the Austric expansion of rice cultivation coming from the south, north of the Yang-Tse Valley. This leads to the conclusion that Austric must have had an ancestral home in Southern China and that Chinese language entered China from the north or northwest. A later intrusion of Chinese into the topographically very broken and fragmented region by political mechanisms would also explain why Austric languages like Miao-Yiao and Austroasiatic are very fragmented in the region (Ruhlen 1988). The incipient Chinese state expands slowly south after 3,000 BC, and overlays the Austric language group called Man in old Chinese sources, Miao-Yiao is a remnant of the aboriginal population of southern China (Chang 1983). The Tibeto-Burmans move into Burma in a process of demic expansion, and the very special topography in this region creates several isolated pockets of Austric. The calculations can be interpreted to suggest that Burushaski gets isolated from other Dene-Sino-Caucasian languages in the Pamir Mountains, close to where it is found today, by Dravidian and Ugrian-Altaic languages around 5,500 BC. The model suggests that Burushaski is closer to the western Dene-Sino-Caucasian language group. Fig. 6 show the relation between the languages included in the Nostratic group, based on the model calculations. Fig. 7 show the relation between the languages included in the Sino-Caucasian group as suggested by the LANGUAGE model.

The domestication of horses in the Central Asian steppe around 4,000 BC lead to development of pastoralism and military superiority by 2,800-2,300 BC. The formation of states starting 2,800 BC and the rise of nomadic pastoralism changes conditions for language dispersal. From now on political factors become more and more important in relation to ecological factors. The LANGUAGE

model does not include such processes, and it cannot predict any of the movements of nomadic peoples nor the effect of policies of the large empires.

## 6 Discussion

The time around 111,000-105,000 BC experienced an interglacial warm period which could have initiated a population diffusion out of Africa into Asia and Southern Europe. The split between Proto-Nostratic and Proto-Sino-Caucasian must have occurred before 17,000 BC, since that is the latest date at which Eskimo-Aleut could have split off from Proto-Nostratic and still made it to America in time. The time around 45,000 BC and alternatively 55,000-62,000 BC experienced global climatic changes that pushed populations around enough to have initiated such a division. Etruscan, Basque, Iberian, Aquitanian, Ligurian and Pict seem to originate in the from the same roots as the Caucasian languages. They must have separated from East Caucasian before 12,000-15,000 BC.

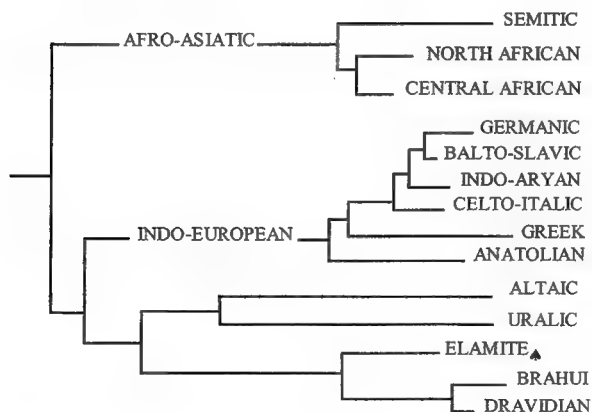


Figure 6. The relation between the languages included in the Nostratic group according to the LANGUAGE model calculation. The model suggest three basic units; Afroasiatic, Indoeuropean and Ugro-Altaic.

The calculations correlate well with the superphylae Nostratic, Sino-Caucasian and Austric. The overall relations between the languages of Eurasia according to the model is shown in Fig. 4. The apparent good fit between language groups and genetically defined populations indicate that initial settlement, demic diffusion and quantitatively significant migrations have been the major mechanisms of language dispersal in Eurasia, and that other mechanisms have not been of any significant importance until 3,000-2,000 BC. The results of the calculations also indicate that glottochronology may significantly underestimate the age of language divergences for dates earlier than 5,000-6,000 BC.

The modified wave-of-advance model employed here seems to be able to describe the present distribution of languages in Eurasia, also on a detailed level in the European and Indian subcontinents. The isolated positions of Basque, Etruscan, Iberian and Pict in Europe are all predicted. Basque could be a relict of the language of the Cro-Magnon man around 35,000 BC, as suggested by Cavalli-Sforza (1989), but there is also a possibility of it being a product of later language migrations connected to the climatic changes at the end of the glaciation in the Gravettian period of 27,000-15,000 BC. The model favors the earlier date for Basque, but the later date for the general spread of Dene-Sino-Caucasian across Eurasia.

The uncertainties in these calculations remain quite large at the present time. Especially the rate coefficients for population growth and migratory diffusivity in mesolithic time remain problematic to estimate. Further research is needed for this type of transfer. The model uses rather large grid cells for Asia at present, and a smaller grid cell would refine resolution as well as precision. Several points in time are possible for the split between Nostratic, Sino-Caucasian and Amerind. Model calculations cannot be sufficiently confined by independent archaeological data and linguistic clues to decide the issue at present.

Modern man expanded into the new world at a rate that implies that there was no population counterpressure. Thus the population density of *Homo Erectus* was either too low to be of consequence, or his competitiveness was so inferior

that the actual population density was of no consequence. The models yield the best results if we simply pretend that nobody was there. Something prevented modern man from entering Europe until 45,000 BC. This could have been the competitiveness of the Neanderthals that was sufficient to prevent modern humans to gain a decisive advantage until 45,000 BC.

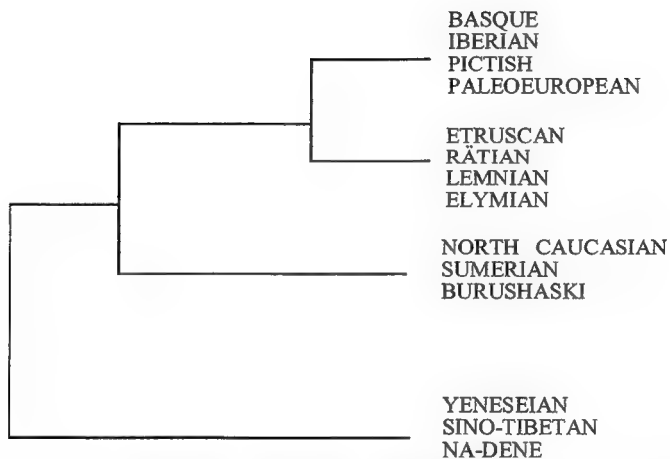


Figure 7. The relation between the languages included in the Sino-Caucasian group as suggested by the LANGUAGE model.

## 7 Conclusions

The LANGUAGE model calculations and the data available indicate a number of statements that may be made:

-Nostratic is supported by the model calculations. The spread of the Nostratic languages can be simulated based on ecological factors and the spread of agriculture 9,000-3,000 BC.

-Dene-Sino-Caucasian and Proto-Nostratic divided before either 35,000, 43,000 or 55,000 BC, caused by ecological changes. The substrate language underlying Indo-European in Europe and Anatolia could have been of Caucasian type. The original language from which the Paleoeuropean languages developed split in the time period 23,000-17,000 BC, when Europe was repopulated after having been a polar desert in a very cold period.

-Austic as a group of genetically related languages is supported by the model calculations. The present distribution of Austic languages can be calculated based on demic diffusion caused by the introduction of rice agriculture around 7,000 BC.

-Amerindian split from Proto-Eurasiatic caused by sufficiently large ecological changes in 22,000, 35,000 or 43,000 BC. Diffusion times required to cover the distance to the Bering Straits make the earlier dates more likely than the late.

The correlation between model calculation and observed genetic pattern Cavalli-Sforza et al, (1978, 1991) is good. The Nostratic language family is supported by both the model and the genetic data.

The compilation of data together with the present calculations, indicate that the major language groupings Nostratic, Sino-Caucasian, Austic and Amerind share a common ancestry lying at least 50,000 years back in the past. Genetic data suggests that modern man emerged from Africa about 100,000 BC. The model requires this amount of time to get every man to his historic position. This seem to indicate that the origin of speech is older than 100,000 years.

LITERATURE

- Allchin, B. and Allchin, R.: 1982, *The rise of civilization in India and Pakistan*, Cambridge University Press, Cambridge.
- Ammerman, A. and Cavalli-Sforza, L.L.: 1971, *Measuring the rate of spread of early farming in Europe*,  
In Man 6, 674-688.
- Ammerman, A. and Cavalli-Sforza, L.L.: 1984, *The neolithic transition and the genetics of populations in Europe*, Princeton University Press, New York.
- Anderson, J.M.: 1988, *Ancient languages of the hispanic peninsula*, University Press of America, New York.
- Anthony, D.: 1986, *The Kurgan culture Indo-European origins and the domestication of the horse: A reconsideration.*,  
In Current Anthropology, 27, 291-314.
- Bellwood, P.: 1989, *The colonization of the Pacific: some current hypotheses*,  
In A. Hill and S. Serjeantson (eds), The colonization of the Pacific, a genetic trail, Clarendon Press, Oxford.
- Bellwood, P.: 1991, *The Austronesian dispersal and origin of languages*,  
In Scientific American, 265, 70-75.
- Bengtson, J.: 1991a, *Macro-Caucasian: A historical linguistics hypothesis*,  
In V. Ševoroškin (ed.), Dene-Sino-Caucasian Languages  
Universitätsverlag Dr. Norbert Brockmeyer, Bochum.
- Bengtson, J.: 1991b, *Macro-Caucasian phonology*,  
In V. Ševoroškin (ed.), Dene-Sino-Caucasian Languages  
Universitätsverlag Dr. Norbert Brockmeyer, Bochum.
- Bengtson, J.: 1991c, *Some Macro-Caucasian etymologies*,  
In V. Ševoroškin (ed.), Dene-Sino-Caucasian Languages  
Universitätsverlag Dr. Norbert Brockmeyer, Bochum.
- Best, J. and Woudhuizen, F.: 1988, *Ancient scripts from Crete and Cyprus*, Brill, Leiden, Netherlands.  
Publications from the Henri Frankfort Foundation 9.
- Bird, R.B., Stewart, W.E. and Lightfoot, E.N.: 1960, *Transport phenomena*, John Wiley and sons, New York.
- Blust, R.: 1988, *The Austronesian homeland--A linguistic perspective*,  
In Asian Perspectives 26, 45-67.

- Bomhard, A. and Kerns, J.: 1994, *The Nostratic macrofamily. A study in distant linguistic relationship*,  
Mouton de Gruyter; Trends in linguistics, studies and monographs 74, Berlin.
- Bonfante, G. and Bonfante, L.: 1983, *The Etruscan language*,  
Manchester University Press, Manchester.
- Bouda, K.: 1938, *Die beziehungen des sumerischen zum baskischen, westkaukasischen und tibetischen*,  
In Mitteilungen der Altorientalische Gesellschaft 12, 1-23.
- Bouda, K.: 1948, *Baskisch und Caucasisch*,  
In Zeitschrift für phonetik und allgemeine sprachwissenschaft, 2, 182-202.
- Broecker, W.C. and Denton, G.H.: 1990, *What drives the glacial cycles ?*,  
In Scientific American 262, 43-50.
- Burenhult, G.: 1981, *Stenåldersbilder-Hällristningar och stenåldersekonomi*,  
Stureförlaget AB, Stockholm.
- Catford, H.: 1991, *The classification of caucasian languages*,  
In S.Lamb and E.D. Mitchell (eds), Sprung from some common source,  
Stanford University Press, Stanford, California.
- Cavalli-Sforza, L.L.: 1988, *The Basque population and ancient migrations in Europe*,  
In Munibe-Antopologia y Arqueologia Suppl. 6, 129-137.  
San Sebastian, Spain.
- Cavalli-Sforza, L.L., Menozzi, P. and Piazza, A.: 1994, *The history and geography of human genes*,  
Princeton University Press, New Jersey.
- Cavalli-Sforza, L.L., Piazza, A., Menozzi, P. and Mountain, J.: 1988, *Reconstruction of human evolution: Bringing together genetic, arcaeological and linguistic data*,  
In Proc. Natl. Acad. Sci. USA 85, 6002-6006.
- Chamorro, J.G.: 1987, *Survey of the archaeological research on Tartessos*,  
In American Journal of Archaeology 91, 197-232.
- Champion, T., Gamble, C., Shennan, S. and Whittle, A.: 1984, *Prehistoric Europe*,  
Academic Press, London.
- Chang, T.: 1983a, *Concluding remarks on the origins of Chinese civilization*,  
In D.N. Keightley (ed.), The origins of chinese civilization, 565-581.  
University of California Press, Berkley.
- Chang, T.: 1983b, *The origins and early cultures of the cereal grains and food legumes*,  
In D.N. Keightley (ed.), The origins of chinese civilization, 65-93.  
University of California Press, Berkley.

- Collinder, B.: 1974, *Indo-Uralish -- oder gar Nostratisch? Vierzig Jahre auf rauhen Pfaden*,  
In G. fuer Hermann Guentert zur 25. Wiederkehr seines todestages (ed.), Antiquitates  
Indogermanicae-Studien zur Indogermanischen altertumskunde und zur sprach und  
kulturgeschichte der indogermanischen völker,  
Innsbruck, pp.363-375.
- dem Elzen, M.: 1994, *Global environmental change, an integrated modelling approach*,  
International Books, Utrecht.
- Diakonoff, I.M.: 1985, *On the original home of the speakers of Indo-European*,  
In Journal of Indo-European Studies 13, 92-174.
- Diakonoff, I.M.: 1990, *Language contacts in the Caucasus and the near east*,  
In T.L. Markey and J.A. Grippin (eds), When worlds collide: The Indo-Europeans and the  
Pre-Indo-Europeans,  
Karoma Publishers Inc., Ann Arbor, pp. 53-66.
- Diakonoff, I.M. and Starostin, S.A.: 1986, *Hurro-Urartian as an Eastern Caucasian language*,  
Muenchener Studien zur sprachwissenschaft.
- Dolgopolsky, A.: 1986, *A probabilistic hypothesis concerning the oldest relationships among the  
language families in Northern Eurasia*,  
In V. Ševoroškin and T. Markey (eds), Typology, Relationship and time,  
Karoma, Ann Arbor.
- Dolgopolsky, A.: 1988, *The Indo-European homeland and the lexical contacts of proto-Indo-  
European with other languages*,  
In Mediterranean Language studies 3, 7-31.
- Doluxanov, P.M.: 1982, *Upper pleistocene and holocene cultures of the Russian plain and  
Caucasus: Ecology, economy and settlement pattern*,  
In Advances in world archaeology 1, 323-358.
- Doluxanov, P.M.: 1986, *Foragers and farmers in west-central Asia*,  
In M. Zvelebil (ed.), Hunters in transition: Mesolithic societies of temperate Eurasia and  
their transformation to farming,  
University of Edinburgh, Edinburgh.
- Fagan, B.: 1992, *The Peoples of the World*  
Harper Collins Publications
- Fairservis, W.A.: 1983, *The script of the Indus valley civilisation*,  
In Scientific American 244, 44-52.



- Formazov, A.A., Černyx, E.N., Šelov, D.B. and Rozenfeldt, R.L.: 1975,  
*The most important recent archaeological discoveries made in European Russia*,  
 In R. Bruce-Mitford (ed.), Recent archaeological excavations in Europe,  
 Routledge and Kegan Paul, London, pp. 188-226.
- Gamkrelidze, T.: 1990, *On the problem of an Asiatic original homeland of the proto-indo-europeans*,  
 In T.L. Markey and J.A. Grippin (eds), When worlds collide: The Indo-Europeans and the Pre-Indo-Europeans,  
 Karoma Publishers Inc., Ann Arbor, pp. 5-14.
- Gimbutas, M.: 1985, *Primary and secondary homeland for the Indo-Europeans*,  
 In Journal of Indo-European Studies 13, 182-202.
- Greenberg, J.: 1957, *Genetic relationships among languages*,  
 University of Chicago Press, Chicago.
- Greenberg, J.: 1987, *Language in America*,  
 Stanford University Press, San Fransisco.
- Greenberg, J.H., Turner, C.G. and Zegura, S.L.: 1986, *The settlement of Americas: A comparison of the linguistic, dental and genetic evidence*,  
 In Current Anthropology 27, 477-497.
- Hewitt, B.: 1981, *Caucasian languages*,  
 In B. Comrie (ed.), Languages of the Soviet Union,  
 Cambridge, Cambridge.
- Hillman, G.C. and Davies, M.S.: 1990, *Measured domestication rates in wild wheats and barley under primitive cultivation, and their archaeological implications*,  
 In Journal of world prehistory 4, 157-222.
- Howells, W.W.: 1983, *Origins of the Chinese people: Interpretations of the recent evidence*,  
 In D.N. Keightley (ed.), The origins of chinese civilization,  
 University of California Press, Berkley.
- Hubschmied, J.: 1960, *Mediterrane substrate*,  
 In Romanica Helvetica 70, 4-97.
- Hubschmied, J.: 1982, *Vorindogermanische und indogermanische substratwörter in den romanischen sprachen*,  
 In S. Ureland (ed.), Die leistung der substratforschung,  
 Max Niemeyer, Tübingen.

- Illič-Svityč, V.M.: 1989, *The relationship of the Nostratic family of languages: A probabilistic evaluation of the similarities question*,  
In V.Ševoroškin (ed.), Explorations in language macrofamilies,  
Universitätsverlag N. Brockmeyer, Bochum, Germany, pp. 111-113.  
Materials from the first International interdisciplinary symposium on language and prehistory at Ann Arbor, 8-12 November, 1988.
- Jackson, K.H.: 1955, *The Pictish language*,  
In F.T. Wainwright (ed.), The problem of the Picts,  
Nelson, London.
- Jönsson, C., Berg, B. and Sverdrup, H.: 1995, *Developing a model of organic matter decomposition based on field and laboratory experiments. 1. The effect of temperature, moisture and acidity*,  
In Reports in environmental engineering and ecology 1:1994, 1-39.
- Kaiser, M. and Ševoroškin, V.: 1987, *On recent comparisons between language families; The case of Indo-European and Afro-Asiatic*,  
In General linguistics 27, 34-46.
- Kammenhuber, A.: 1975, *The linguistic situation of the 2nd millennium BC in ancient Anatolia*,  
In Journal of the Royal Asiatic Society 87, 116-120.
- Kretschmer, P.: 1943, *Die Vorgriechischen Sprach und Volksschichten*,  
In Glotta 30, 84-217.
- Kuper, R.: 1979, *Vom Jäger zum Hirten—Was ist das Sahara-Neolithicum ?*,  
Sahara; 10000 Jahre zwischen Weide und Wüste, Museen der Stadt Köln, Köln.
- Lafon, R.: 1951, *Concordances morphologiques entre le Basque et les langues Caucasiques*,  
In Word 7, 227-244, 8, 80-94.
- Lamberg-Karlovsky, C.-C.: 1988, *The proto-Elamites on the Iranian plateau*,  
In Journal of world prehistory 2, 359-396.
- Laričev V., Xoluškin U., Laričeva I.: 1988, *The upper paleolithic of northern Asia; Achievements, problems and perspectives*  
In Journal of World Prehistory 2, 359-396.
- Ledyard, G.: 1975, *Galloping along with the horseriders, looking for the founders of Japan*,  
In Journal of Japanese studies 1, 217-254.
- Lewin, B.: 1976, *Japanese and Korean: the problems and history of a linguistic comparison*,  
In Journal of Japanese studies 2, 389-412.
- Locker, E.: 1962, *Die ältesten sprachschichten westeuropas*,  
In Österreichische akademie der wissenschaften, philosophische-historische klasse, sitzungsberichte 240, 5-59.

- Mallory, J.P.: 1989, *In search of the Indoeuropeans; Language, archaeology and myth*, Thames and Hudson, London.
- Matyushin, G.: 1986, *The mesolithic and neolithic in the southern Urals and central Asia*, In M. Zvelebil (ed.), Hunters in transition: Mesolithic societies of temperate Eurasia and their transformation to farming, University of Edinburgh, Edinburgh.
- McAlpin, D.W.: 1981, *Proto-Elamo-Dravidian: The evidence and its implications*, In Transactions of the American Philosophical Society 71:3, 1-155.
- Meacham, W.: 1983, *Origins and development of the Yüeh coastal neolithic: A microcosm of cultural change on the mainland of east Asia*, In D.N. Keightley (ed.), The origins of chinese civilization, University of California Press, Berkley.
- Menges, K.: 1989, *East-nostratic: Altaic and Dravidian*, In V. Ševoroškin (ed.), Proto-Languages and Proto-Cultures, Universitätsverlag N. Brockmeyer, Bochum, Germany, pp. 59-62.  
Materials from the first International interdisciplinary symposium on language and prehistory at Ann Arbor, 8-12 November, 1988.
- Menges, K.: 1990, *Altaic and east Nostratic*, In V. Ševoroškin (ed.), Proto-Languages and Proto-Cultures, Universitätsverlag N. Brockmeyer, Bochum, Germany, pp. 26-32.  
Materials from the first International interdisciplinary symposium on language and prehistory at Ann Arbor, 8-12 November, 1988.
- Mukarovsky, H.G.: 1969, *Baskisch-berbische entsprechenungen*, In Wiener Zeitschrift für die Kunde des morgenlandes 62, 35-51.
- Nelson, S.: 1990, *The neolithic of northeastern China and Korea*, In Antiquity 64, 234-248.
- Nikolaev, S.: 1991, *Sino-Caucasian languages in America*, In V. Ševoroškin (ed.), Dene-Sino-Caucasian Languages, Universitätsverlag Dr. Norbert Brockmeyer, Bochum.
- Nikolaev, S. and Mudrak, O.: 1989, *Gilyak and Chukchi-Kamchatkan as Almosan-Keresiouan languages; lexical evidence*, In V. Ševoroškin (ed.), Explorations in language macrofamilies, Universitätsverlag N. Brockmeyer, Bochum, Germany, pp. 67-89.  
Materials from the first International interdisciplinary symposium on language and prehistory at Ann Arbor, 8-12 November, 1988.

- Nikolaev, S. and Starostin, S.: 1991, *North Caucasian roots*,  
In V. Ševoroškin (ed.), Dene-Sino-Caucasian Languages  
Universitätsverlag Dr. Norbert Brockmeyer, Bochum.
- Nissen, H.J.: 1983, *The early history of the ancient near east 900-2000 BC*,  
University of Chicago Press, Chicago.
- Orël, V. and Starostin, S.: 1990, *Etruscan as an East-Caucasian language*,  
In V. Ševoroškin (ed.), Proto-Languages and Proto-Cultures,  
Universitätsverlag N. Brockmeyer, Bochum, Germany, pp. 60-67.  
Materials from the first International interdisciplinary symposium on language and  
prehistory at Ann Arbor, 8-12 November, 1988.
- Pallotino, M.: 1975, *The Etruscans*,  
Allen Lane, Penguin Books, London.
- Pearson, R. and Lo, S.: 1983, *The Ch'ing-lien-kang cultures and the Chinese neolithic*,  
In D.N. Keightley (ed.), The origins of chinese civilization,  
University of California Press, Berkley.
- Peiros, I.: 1990, *The ancient eastern and southeastern Asia: Comparative-historical data and their  
interpretation*,  
In V. Ševoroškin (ed.), Proto-Languages and Proto-Cultures,  
Universitätsverlag N. Brockmeyer, Bochum, Germany, pp. 14-25.  
Materials from the first International interdisciplinary symposium on language and  
prehistory at Ann Arbor, 8-12 November, 1988.
- Peiros, I.A. and Starostin, S.A.: 1984, *Sino-Tibetan and Austro-Tai*,  
In Computational analysis of Asian and African languages 22, 123-127.
- Peiros, I. and Shnirelman, V.: 1989, *Toward an understanding of proto-dravidian prehistory*,  
In V. Ševoroškin (ed.), Proto-Languages and Proto-Cultures,  
Universitätsverlag N. Brockmeyer, Bochum, Germany, pp. 70-73.  
Materials from the first International interdisciplinary symposium on language and  
prehistory at Ann Arbor, 8-12 November, 1988.
- Pope, G., Barr, S., Macdonald, A. and Nakabanlang, S.: 1986, *Earliest radiometrically dated  
artifacts from the southeast Asia*,  
In Curent Anthropology 27, 275-279.
- Pulleyblank, E.G.: 1983, *The Chinese and their neighbors in prehistory and early historic times*,  
In D.N. Keightley (ed.), The origins of chinese civilization,  
University of California Press, Berkley.
- Reid, L.: 1988, *Benedict's Austro-Tai hypothesis--an evaluation*,  
In Asian Perspectives 26, 19-34.

- Renfrew, C.: 1989a, *Archaeology and language-The puzzle of Indo-European origins*, Penguin Books, 27 Wrights Lane, London W8 5TZ, Great Britain.
- Renfrew, C.: 1989b, *Models of change in language and archaeology*, In *Transactions of the Philological Society* 87, 103-155.
- Renfrew, C.: 1994a, *Before Babel, Speculations on the origins of linguistic diversity*, In *Cambridge Archaeological Journal* 1, 3-23.
- Rhys, J.: 1892, *The inscriptions and language of the northern Picts*, In *Proceedings of the Scottish Archaeological Society* 26, 263-351.
- Ruhlen, M.: 1988, *Guide to the world's languages*, Stanford University Press, Stanford.
- Ruhlen, M.: 1985, *Nostratic-Amerind cognates*, In V. Ševoroškin (ed.), *Proto-Languages and Proto-Cultures*, Universitätsverlag N. Brockmeyer, Bochum, Germany, pp. 75-83.  
Materials from the first International interdisciplinary symposium on language and prehistory at Ann Arbor, 8-12 November, 1988.
- Scharf, J.: 1980, *Die sapienten-population im neolithicum zentral und nordeuropas-hypothesen, model und realität*, In *Gegebaurs morph jahrbuch* 126, 449-477.
- Schmidt, K.H.: 1985, *A contribution to the identification of Lusitanian*, In J.Hoz (ed.), *Actas del III coloquio sobre lenguas y culturas paleohispanicas*, Universidad de Salamanca, Spain.
- Schmoll, U.: 1961, *Die Südlusitanischen Inschriften*, Harassowitz, Wiesbaden.
- Sherratt, A. and Sherratt, S.: 1988, *The archaeology of Indo-Europe; An alternative view*, In *Antiquity* 62, 584-595.
- Ševoroškin, V. and Manaster Ramer, A.: 1991, *Some recent work on the remote relations of languages*, In S. Lamb and E.D. Mitchell (eds), *Sprung from some common source*, Stanford University Press, Stanford, California.
- Starostin, S.: 1989, *Nostratic and Sino-Caucasian*, In V. Ševoroškin (ed.), *Explorations in language macrofamilies*, Universitätsverlag N. Brockmeyer, Bochum, Germany, pp. 42-66.  
Materials from the first International interdisciplinary symposium on language and prehistory at Ann Arbor, 8-12 November, 1988.

- Starostin, S.: 1991, *On the hypothesis of a genetic connection between the Sino-Tibetan languages and the Yenesseian and North Caucasian languages*,  
In V. Ševoroškin (ed.), Dene-Sino-Caucasian Languages  
Universitätsverlag Dr. Norbert Brockmeyer, Bochum.
- Stringer, C.: 1990, *The emergence of modern humans*,  
In Scientific American 263, 68-74.
- Sverdrup, H.: 1995, *Classifying and translating inscriptions in the Pictish language*,  
In Reports in ecology and environmental engineering 1995:4.
- Sverdrup, H.: 1998, *A typological and morpho-syntactical analysis of the extinct Rätian language*,  
In Reports in ecology and environmental engineering 1997:3.
- Sverdrup, H.: 1998, *Ecological modelling of language origins, relatedness and paths of dispersal; Part 1; The LANGUAGE model*.  
To be published in the series Reports in ecology and environmental engineering during 1998.
- Sverdrup, H.: 1998, *Ecological modelling of language origins, relatedness and parths of dispersal; Part 2: Linguistic interpretation of results from the LANGUAGE model*.  
To be published in the series Reports in ecology and environmental engineering during 1998.
- Sverdrup, H. and Guardans, R.: 1998a, *The Swedish-Spanish Iberian language study 1: Analysis of some utilitarian texts in the ancient Iberian language*.  
Manuscript prepared for submission during 1999
- Sverdrup, H. and Guardans, R.: 1998b, *The Swedish-Spanish Iberian language study 2: Analysis and attempted interpretation of the lead from Alcoy*.  
Manuscript prepared for submission during 1999
- Sverdrup, H. and Guardans, R.: 1998c, *The Swedish-Spanish Iberian language study 3: A morpho-syntactical study of texts from Solaig, Emporion, Castellon and Ullastret*.  
Manuscript prepared for submission during 1999
- Sverdrup, H. and Guardans, R.: 1998d, *The Swedish-Spanish Iberian language study 4: Genetic relationships and typology of the ancient Iberian language*.  
Manuscript prepared for submission during 1999
- Tovar, A.: 1977, *Krahes alteuropäische hydronomie und die westindogermanische sprachen*,  
Carl Winter Universitätsverlag, Heidelberg.
- Turner, C.: 1987, *Late pleistocene and holocene population history of East Asia based on dental variation*,  
In American Journal of Physical Anthropology 73, 305-321.

- Untermann, J.: 1990, *Monumenta Linguarum Hispanicarum. Band III; Die Ibersichen Inschriften aus Spanien –a) Literaturverzeichnis, Einleitung, Indices, –b) Die Inschriften*, Dr. Ludwig Reichert Verlag, Wiesbaden.
- Vennemann, T.: 1994, *Linguistic reconstruction in the context of European prehistory*, In Transactions of the Philological Society 92, 215-284.
- Vetter, E.: 1943, *Die übrigen vorrömischen sprachen im Oberitalien*, In Glotta 30, 66-81.
- Villar, F.: 1990, *Indo-europeéns et pre-indo-europeéns dans la péninsule Ibérique*, In T.L. Markey and J.A. Grippin (eds), When worlds collide: The Indo-Europeans and the Pre-Indo-Europeans, Karoma Publishers Inc., Ann Arbor, 363-394.
- Woudhuizen, F.: 1992a, *The language of the sea peoples*, Najade Press, Amsterdam, Netherlands.  
Publications from the Henri Frankfort Foundation 12.
- Zohary, D.: 1990, *Domestication of plants in the old world; The emerging synthesis*, In T.L. Markey and J.A. Grippin (eds), When worlds collide: The Indo-Europeans and the Pre-Indo-Europeans, Karoma Publishers Inc., Ann Arbor, pp. 35-44.
- Zvelebil, M.: 1986, *Mesolithic societies and the transition to farming; problems of time, scale and organization*, In M. Zvelebil (ed.), Hunters in transition: Mesolithic societies of temperate Eurasia and their transformation to farming, University of Edinburgh, Edinburgh.





# COMPILING WORDS FROM EXTINCT NON-INDOEUROPEAN LANGUAGES IN EUROPE

Harald Sverdrup and Ramon Guardans

## 1 Introduction

In the past we have studied several of the extinct Paleoeuropean languages of Europe, such as Pictish (Sverdrup 1995), Rätian (Sverdrup 1997), Tartessian (Sverdrup and Guardans 1999) and Iberian (Sverdrup and Guardans 1998; 1999a,b,c). We have assembled a database on words from prehistoric European languages to use in our studies. The Indo-European languages were not alone in Europe in antiquity, nor are they today. Basque is the only surviving language of a once larger group of languages in Europe, which were spoken in Europe before agriculture. In the past there was an unknown multitude of now extinct languages in Europe that evolved over space and time in a complex web of relations. Currently we can only deal with the simplified image of this diversity as preserved in a few remains beside the living Basque language. We have documented words from these languages as far as we are able to find them. Glossaries have been collected for Camunian, Elymian, Etruscan, Iberian, Lemno-Pelasgian, Lepontic, Ligurian, North Pikene, Nuragic, Northwestern Paleoeuropean, Itturan, Southeastern Paleoeuropean, Pictish, Rätian, Tartessian. We will call these languages Paleoeuropean, a geographical grouping at the moment. Extinctions can be associated with the spread of Indo-European languages and the neolithic revolution, the expansion of Greek and Roman empires and the collapse of the Roman empire. Aquitanian has been shown to be probably ancestral to Basque, and is not included in this study.

## 2 Objective and scope

The purpose of this study is to make our compilation of glosses from extinct Paleoeuropean languages available to a wider audience. We find it important that others have a chance to inspect our material. This study does not pretend to present any in-depth analysis of word etymologies or propose any genetic relationships. The fact that there are at least 16 different identifiable relict languages of probable or certain non-indoeuropean character in Europe, invite us to compare whatever is left of these languages with each other, regardless of whether they are genetically related or not. However, the fact that there are so many Paleoeuropean languages makes it more likely that at least some of them are related. The scope of this work is limited to Europe and the survey of glosses from now extinct non-Indoeuropean languages.

## 3 Data

The data was taken from a number of sources: Pictish (Rhys 1892, 1898; MacAllister 1940; Jackson 1955; Sverdrup 1995); Iberian (Anderson 1988; Pattison 1981; Untermann 1975, 1980, 1990a,b; Sverdrup and Guardans 1999a,b,c); Tartessian (Villar 1990, Sverdrup and Guardans 1999d); Rätian (Schumacher 1992, Sverdrup 1997). Etruscan (Pallotino 1975, Cristofani 1979, Rix 1968, 1985, 1991; Bonfante and Bonfante 1983 and Bonfante 1990). Further sources used are Kretschmer (1942, 1943), Ambrosini (1979), Hubschmied (1960, 1982), Schuhmacher (1992), Vennemann (1994).

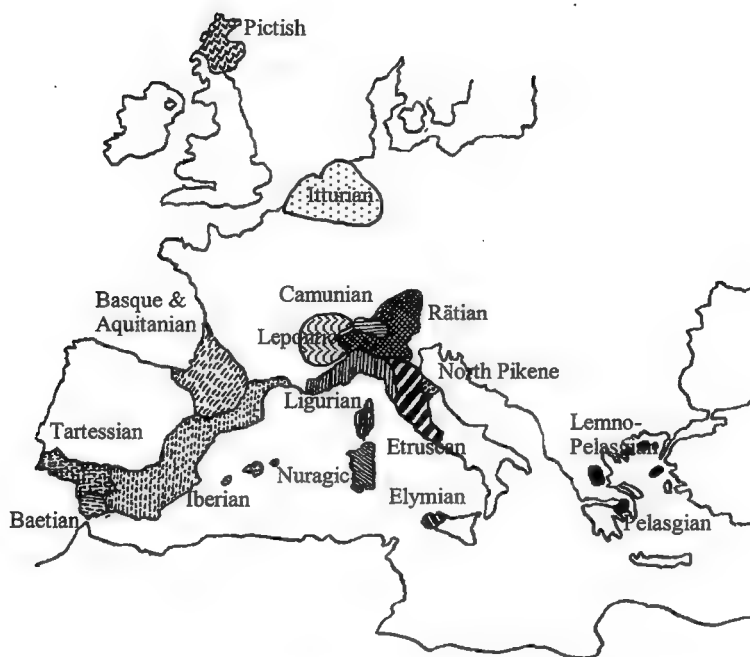


Figure 1. *The approximate position of known Paleoeuropean languages in approximately 1,000-500 BC.*

Camunian is only attested from the graffiti on the rock carvings of Val Camonica in northern Italy. No longer texts are known, the longest text contain maybe 5 words. The language in the inscriptions is different from the Rätian language found in adjacent areas, some believe it to be of Celtic derivation. No systematic studies of the language has been published.

Elymian existed in antiquity in the westernmost tip of Sicily in the cities of Eryx and Segesta. It disappeared shortly after the Roman conquest of the island. Only fragmentary inscriptions on broken pottery are known.

Etruscan existed in Italy until late Roman times. The language is well attested in many inscriptions, of which many are long. Inscriptions can generally be understood in broad outline, but the details of the language are poorly understood. Of approximately 8,000 glosses known, we have been able to find translations for approximately 550 glosses. There is a very large literature available on Etruscan morphology, phonology, grammar and genetic relationships.

Iberian is the indigeneous language of southeastern Hispanic peninsula. Approximately 3,500 texts are available, several of these are quite long. Nothing of the language is understood and hardly a gloss can be translated, according to the current learned opinion. The authoritative documentation on Iberian is the large corpus of Iberian inscriptions by Prof. J. Untermann of Köln University (*Monumenta Linguarum Hispanicum*, Vols. I-III, Untermann 1975, 1980, 1990a,b). Recent work by Roman del Cerro (1993) and the Swedish-Spanish language study (Sverdrup and Guardans 1999a,b,c), have yielded a number of translations which have been listed below.

Lemnno-Pelasgian is made up of the language from the Kamnia stela and potshards from the Greek island Lemnos and glosses recorded in antique manuscripts as Pelasgian. The language appears to be closely related to Etruscan (Rix 1968), and the available material is sufficient to firmly support this thesis. It disappeared in antiquity.

Lepontic appears in the border region between Switzerland and Italy along the large lakes. It is attested in a small number of inscriptions from gravestones, too little for certain classification. Current theories has it that it is either Celtic, Gallic over a non-Indoeuropean substratum or Etruscoid.

Ligurian was the indigeneous language of the Ligurian Mountains and Riviera. It went extinct during Roman times, and almost no inscriptions have been

preserved. Many substrate words in the area show similarity to Basque, but far too little remains for any certain classification.

Nuragic was the autochthonous language of Sardinia. The language is mainly known from substrate studies and placenames. Many substrate words in the area show similarity to Basque, but far too little remain for any certain classification.

North Pikene is basically only known from the Novilara stele in middle eastern Italy. The genetic affiliation of the language on this single stone is very uncertain, except for concluding that the language is probably not of Indoeuropean type.

Paleoeuropean is the first language of early post-glacial Europe. Possibly, there may have been a Northwestern and a Central branch. It can be reconstructed from placenames and substrate studies and shows significant similarity to Basque (Vennemann 1994). Northwestern Paleoeuropean was covering most of western, central and eastern Europe, and Central Paleoeuropean was covering all Italy and the Balkan side of the Adriatic coast. Itturan was a branch of Northwestern Paleoeuropean that survived until approximately 800 AD (Vennemann 1994, Locker 1962). The languages are only attested from toponyms.

Pictish was the autochthonous language of Scotland until approximately 850 AD (Rhys 1892; MacAllister 1940). It is recorded in Ogham letters on stone monuments. Approximately 30 inscriptions have been preserved. Current theories being seriously discussed are if it is of Celtic affiliation or of non-Indoeuropean derivation. Two studies give a full corpus and discuss affiliation (Sverdrup 1995; Forsyth 1996).

Rätian is a language attested on artefacts and monuments from Tirol and Northern Italy. Approximately 50 inscriptions are known, some of them are substantial. The current theory is that the language is affiliated with Etruscan (Schumacher 1994; Sverdrup 1997).

Tartessian is a language attested from the southwestern part of the Iberian peninsula. Until the present the inscriptions could not be read, but a recent study has

proposed a system for transcription, and a corpus has been made (Sverdrup and Guardans 1999d). Very few words can be translated, the genetic affiliation is unknown.

Language	Number of glosses translated	Approximate total number of glosses	% of known vocabulary translated
Etruscan	665	4000	16
Iberian	180	2300	8
Rätian	101	185	50
Pictish	92	104	88
Tartessian	9	85	11
Northwestern Paleoeuropean	30	50	67
Central Paleoeuropean	4	15	25
Camunian	7	46	15
Lepontic	7	41	17
Elymian	26	40	72
North Pikene	5	39	13
Lemnian	26	40	72
Nuragic	14	40	25
Ligurian	5	16	30
Itturian	6	8	70
Pelasgian	9	20	50

Table 1. *Statistics for the collected languages. The first column is the number of glosses that can be translated or a probable meaning assigned, excluding names.*

Some statistics of the glossaries found for 14 languages have been shown in Tab.1. For Etruscan, a larger part of the vocabulary is probably known. Rix (1991) gives a vocabulary list of approximately 8,250 items, but many of these are names and repetitions of similar forms and spellings of identical words. A rough estimate would give that approximately 4,000 different glosses are known in Etruscan. Sometimes it

is said that "Etruscan has not yet been deciphered", apparently irritating for those studying Etruscan. It can be seen from Tab. 1 that Etruscan is only partially understood, only a mere 14% of all the known words are understood. So it appears that quantitatively, Etruscan has only been partially deciphered. But this does not tell the whole story, as the numbers do not contain the quality of the Etruscan translations. Many Etruscan texts are actually quite well understood, the decipherments have been quite successful, even if parts of the grammar seems obscure. But it can also not be denied that certain very important words are not known (foot, eye, hair, you, they, uncle, arm, leg, no, yes, blood, food, grass, cow, pig, make, run, walk, etc.). Etruscan is far better understood than Rätian, even if a larger part of the smaller Rätian vocabulary can be translated.

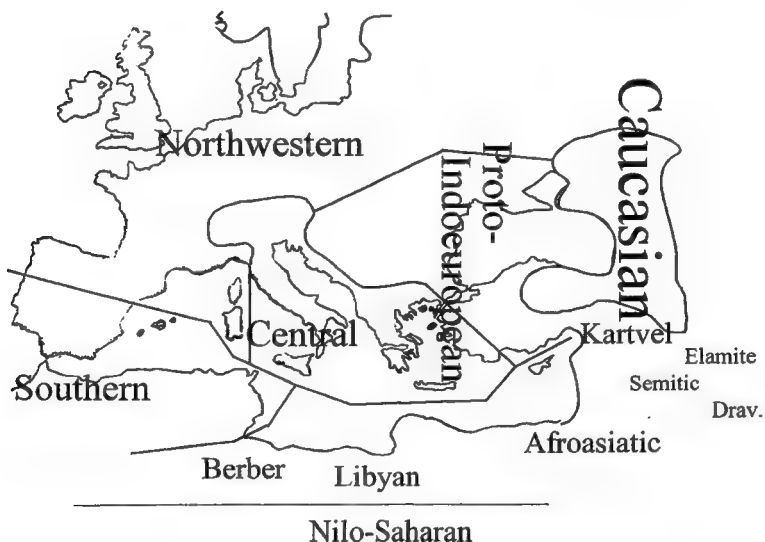


Figure 2. *The approximate position in 5,500 BC of the major Paleoeuropean groups in Europe, based on similarities discovered in the glossaries and Krahe's old European hydronymy (Vennemann 1994). Paleoeuropean was possibly divided internally into Southwestern, Northwestern, Central and Caucasian. The area indicated as Proto-Indoeuropean was probably at this time beginning to be occupied by advancing neolithic farming peoples. In this area, no toponyms characteristic of the old European hydronymy can be found. South of the Mediterranean, Afroasiatic languages advanced as a result of the neolithic transition*

For Iberian, as many as 2,000 words are probably known, but only a small amount of these have been translated. Still researchers disagree over every new approach, demanding absolute proof. In light of the total lack of success of this strategy for interpreting Iberian texts, it goes without saying that new methods and allowance for more innovation and fantasy must be made in order to make progress. The amount of words and texts available are comparable to the volume of Etruscan texts, and the long term prospect for deciphering Iberian should be good.

Lemnian and Rätian can partly be understood in broad terms because of similarity to Etruscan. At present Elymian also appear to be similar to Etruscan, but too little is left for a real assessment. The interpretation of Pictish glosses was assisted by archaeological context as well as by similarity to Basque, but too little is left to prove anything, nor make a genetic relationship certain. For the other Paleoeuropean languages we can only make conjectures and show similarities and correlations, but the material is too small to prove anything. The inscriptions on the beaked vessel from Castaneda in the Alps between the Lepontic and the Camunian area have been added to the Camunian glossary. The inscription from San Bernardino has been added to the Lepontic glossary. The glossaries have been listed according to language in alphabetical order. Where we have noticed that the word seems similar to that of another language, we have commented on this. The similarities are not made with any in-depth etymologizing or insight; the method is simple taxonomy. Thus several similarities may be spurious. The meanings of many of these words are



unknown, but we have listed the meanings or guessed meanings as far as we have been able to find suggestions or established meanings. A questionmark will identify uncertain meanings of a particular word. In the glossaries, the following notation has been used: <sup>+</sup> is a word attested in manuscripts in Latin or Greek from antiquity, \* is a reconstructed etymology and <sup>s</sup> is a word from substrate studies and toponyms. We have mainly used published sources, but also personal communications, institute reports and unpublished manuscripts.

#### 4 The western group

It is possible to see a western group of languages in paleolithic Europe. This group comprise Basque, Pictish and Iberian. These languages share some grammatical traits and have similar words. Pictish, Basque and Iberian also appear to have several numeral words in common. The available material is suggestive of a genetic relationship (Sverdrup 1995; Sverdrup and Guardans 1999a,b,c), but insufficient in volume to yield real proof. In the area between the rivers Maas and Aller in the Netherlands, remnants of a non-Indoeuropean language which we will call Itturian, has been identified. Itturian is documented only from placenames. Placenames and the word-stems have been compared to Basque (Vennemann 1994).

Pronominal	Iberian	Elymian	Pictish	Rätian	Etruscan	Basque	Caucasian
							ze
1. sg. I	nai			nai	-ni	*nai	ni
	mi	mi	mi	mi	mi		mm
2. sg.thou	gu		gu	(ku)	χu	*ki	gu
3. sg. it	te/-e				-ne	e/a	
1. pl. we	(iu)				(ti)	gu	zi, ti
2. pl. you	su			(su)	(su)	zu	zu, wo
3. pl. they	?				(nu)	ak	

Table 2. *Some of the established or proposed pronouns from the glossary as compared to Basque. Uncertain determinations are marked within brackets ().*

One example is *itter* as a stem in place-names connected with wells and springs, similar to the Basque word *iturri*, meaning "spring, well". Of the Ligurian and Nuragic languages very little material remain, Hubschmied (1960, 1982) suggests that the few substrate words available have similarities with Basque.

#### The central group

The central group consists of the Etruscan, Lemnian and Rätian languages which appear to be genetically related (Schuhmacher 1992, Rix 1968, Sverdrup 1997). These languages share important grammatical, morphological and phonetic traits, making a relationship certain. The Lemnian language is recorded as the language of Samothrace, Imbros, Thasos and Lemnos in the Aegean Sea. The position of Lemnian in the Aegean Sea and the Greek mainland together with Etruscan and Rätian in the Alps, would be explained if it was assumed that these languages are remnants of the pre-neolithic languages of Europe. In Etruscan, we have identified occurrence of prefixes on nouns. Etruscan is an agglutinative language, excluding it from Indoeuropean. Rätian is agglutinative, and it is possible to show the remnants of a prefix system in it. The position of the Elymian language is much harder to assess beyond formulating a hypothesis. Very little of the language is preserved, mostly fragments of words. Enough of the vocabulary is preserved to suggest that the language was related to Etruscan.

#### 6 Uncertain affiliations

The North Pikene language is basically untranslated. The glossary contain a few words reminding of Etruscan, the suffix structure point towards an agglutinating structure, excluding it from being Indoeuropean.

The Camunian language shows both Indoeuropean and Etruscoid features. Suffixes like *-iz*, *-az*, *-us* may appear Indoeuropean, but suffixes like *-ai*, *-al*, *-ti*, *-pi* and *-si* follow the Paleo-Mediterranean pattern of non-Indoeuropean character, also

found in Etruscan. The apparent use of prefixes and agglutination of postpositions suggest a non-Indo-European origin.

Value Etr. Bsq. Iberian Rät. Pict PaleoE Lemnian Caucasian

1	(-pa)	bat	ba	(-pa)	bant	bant		
	thu						zu	
		ike					ake	
2		bir	bi		(ir)			
	zal	*zor	(*sor)				si	
3	ki	hir	(kiao)	ki			ki	
4	huth	lau	(lu)			hytt	-omk-	
5		bortz					pchi	
	mak		m(ek)	m..		maras	mki	
6	sa	sei		(sei)		ne		
7	kezp	zazpi						
8	semph	zorzi	(sorze)					
9	nurph	bed-era-zi						
10	zar	*zi	se	(s <sup>u</sup> i)		sar	swi	
100	tha	ka	ta				tarsh	

Table 3. Comparison of words interpreted as numerals, found in the glossaries, together with Basque. Very uncertain determinations are given in brackets ( ). For several of the languages in the glossary, no numeral words can be securely identified

Lepontic has been classified as both Liguroid and Indo-European in the literature. The Lepontic personal names appear to have Gallic character and bear Gallic suffixes (-os, -ui, -om, -ikna, -a). But the few words and some of the possessive suffixes actually preserved, are not Indo-European and have more of a Etruscan-Rätian flavor (-na, -al, -ti-, -ai, -a). The language appear to agglutinate suffixes, a feature foreign to Indo-European. Guessing to a large degree, our feeling is

that the Gallic element in Lepontic was caused by elite dominance over a non-Indoeuropean substratum.

Tartessian is a different language in the Iberian peninsula. It shows certain similarities with Iberian and Basque in the suffixes and structure, but nothing is really understood. The large number of different signs for k, t and s may reflect that the language use a large variety of consonants, something typical of Caucasian languages. Tartessian shows none of the nasal phonemes (nb, nd, ng, mb, mp, md, mn) found in many African languages and which also occur to some extent in Iberian and Basque. Toponyms in the area show similarities and characteristics of toponyms in northern Spain and with Basque as well as the central European toponymy (Tovar 1977, Vennemann 1994). This toponymy continues into Morocco and the Atlas mountains (Roman del Cerro 1993) or even into the Canary Islands (Scharf 1978). Thus, a southern Paleoeuropean branch cannot be fully excluded.

Camunian			munthal			
Elymian		athsena	mutai	mlak	iru	tulai
Etruscan		athre	munth	mlake	spura	tular
Iberian		es	mun	mliki	ir	tolir
Itturian						itur
Lemnian						itter
Lepontic						
Ligurian		at				
North Pikene		aiten				
Nuragic						
NW Paleoeuropean			mund			itter
C Paleoeuropean			munth			ittur
Pelasgian		aito			tuli	
Pictish			mundnaith		toles	
Rätian		es		melka	tulie	
Tartessian		eteri	mundu		spura	
Basque		etche	mun	meliska	iro	
Conceptual		house	land,	beautiful	iru	itorri
meaning		hut	world	lick, mouth	city	spring
				tribe	border	drinkingwater

Camunian					asaz	
Elymian		uakar				
Etruscan		ukar	ceri	suthi	esca	ais, aiser
Iberian		urke	karri	seltar	isker	as
Itturian		karra				purthne
Lemnian		aker				bor
Lepontic				sutai	eiser	
Ligurian			karra			
North Pikene				sut, sot-er		
Nuragic		urgu	karr	sati	iska	aesar
NW Paleoeuropean			karr			
C Paleoeuropean			karra			
Pelasgian						prut
Pictish			ceri			brude
Rätian			karr		isker	aiser
Tartessian		urkoe		suduu		burus
Basque		urke	karri	seldor	ezker	oz
Conceptual		high	stone	grave	hand	God
meaning			hell		power	sky
						leader

Table 4. A comparison of some of the words found, the proposed meaning of the original etymology in approximation, all compared to Basque.

Camunian		pali				
Elymian			kis	mi		
Etruscan	laucumo		peras	mi	ni	
Iberian	laku	bala	(i)ber	gis	mi	nai
Itturan		pala				
Lemnian		pella				
Lepontic		pala				
Ligurian		pali	bero			
North Pikene	lakut				nis	
Nuragic	lakon		bero	ni	(nec)	
NW Paleoeuropean						
C Paleoeuropean						
Pelasgian	lake					
Pictish	lugh		perr	gis	me	
Rätian	e-luku	pele	bera		mi	nai
Tartessian						
Basque	lagun	baltsa	beratz	gizon		ni
Conceptual	man,	hollow,	rich,	man,	me,	I,
meaning	chieftain	cave	plenty	male	myself	self
.						
Camunian						
Elymian						
Etruscan	paite	clan	mechlum		undi	tiu
Iberian	baite	kalun			ande	tiu
Itturan					andr	is
Lemnian						
Lepontic			metel			
Ligurian	baite					
North Pikene					diu	
Nuragic						is
NW Paleoeuropean			med		andr	is
C Paleoeuropean			meth		ander	is
Pelasgian	baite					is
Pictish		clar			onde	
Rätian	baida	kalun	meclo		anto	is, isarki
Tartessian		(kontei)				tioo
Basque	baides	kalun			andia	iz
Conceptual	house,	son,	community,	large	moon,	water,
meaning	at home	male	many		month	stream

Table 5: A comparison of some of the words found, the proposed meaning of the original etymology in approximation, all compared to Basque.

Camunian				
Elymian				
Etruscan	(sal)		anai	laur
Iberian	sal	ibar	anaio	lu, laur
Iturian	sal			
Lemnian				
Lepontic				
Ligurian			anaies	
North Pikene				
Nuragic				
NW Paleoeuropean	sal	bar		laur
C Paleoeuropean	sal	ibar		
Pelasgian				laur
Pictish				
Rätian	(sil)			lur
Tartessian		ibun		
Basque	sil	ibar	anai	laur
Conceptual meaning	river, water	hand, possession	brother, numeral 4	land, soil

Table 6: *A comparison of some of the words found, the proposed meaning of the original etymology in approximation, all compared to Basque. We have also listed consonants omitted in the languages and the use of prefixes in the languages.*

	Excluded sounds	use of prefixes in grammar	Ergativity
Camunian	b	yes	no
Elymian	b	?	?
Etruscan	b, o	yes	no
Iberian	f, v	yes	yes
Iturian	f	yes	?
Lemnian	b	?	no
Lepontic	f	?	no

Ligurian	f	?	?
North Pikene	f	no	yes ?
Nuragic	-	?	?
NW Paleoeuropean	f	yes	?
C Paleoeuropean	-	yes ?	?
Pictish	f	yes ?	yes
Rätian	b	yes	no
Tartessian	f, v	no ?	no ?
Basque	f, v, (m)	yes	yes

Table 7: *Consonants omitted in the languages and the use of prefixes in the languages.*

## 7 Discussion

Fig. 2 shows the approximate position in 5,500 BC of the non-Indoeuropean language groups, based on similarities discovered in the glossaries and in Krahe's old European hydronymy (Tovar 1977; Venneman 1994). The Paleoeuropean languages advanced along the retreat of the ice after the last glaciation. Until 12,600 BC and 10,200-9,800 (younger dryas period) the area north of the Alps was cold and arid tundra. The areas marked Proto-Indoeuropean was probably at this time beginning to be occupied by neolithic farming peoples with proto-Indoeuropean languages. In this area no toponyms contain the old European hydronymy. The pattern seen in the map would suggest four groups of Paleoeuropean languages in Europe, firstly comprising the old languages in Northern Spain and northern Europe, secondly in the southern Iberian peninsula, thirdly in the central Mediterranean-Aegean area and fourth in the Anatolia-Caucasus area. These areas were also possibly the foci of the secondary demic expansion into central and northern Europe when the ice retreated at the end of the last ice-age. Table 2 shows pronouns from some Paleoeuropean languages as far as we have been able to establish them. Table 3 shows Paleoeuropean numerals as far as we can establish them, for several of them, the numeral words are not available.



Tables 5 and 7 show some of the similarities found among the glossaries, as well as also a list of prefix use (potential for noun class system) and consonants omitted from the language. Systematically, some of the languages omit *b* (Camunian, Elymian, Etruscan, Lemnian, Lepontic), some omit *f* (Iberian, Itturian, Ligurian, North Pikene, Northwestern Paleoeuropean, Tartessian and Basque). We think the Paleoeuropean languages are remnants of the languages of Europe before the Indoeuropean languages. Our opinion is that Basque belongs to a Paleoeuropean language grouping of at least 4 related languages (Aquitanian, Basque, Iberian, Pict), but maybe as many as 7 (Aquitanian, Basque, Iberian, Pict, Ligurian, Nuragic, Itturian). Northwestern Paleoeuropean was the ancestor of this whole group, including Basque, and it is preserved in toponyms and some substrate words. There are indications that the ancestral language covered the whole of Europe (Fig. 2, see also Vennemann 1994). Etruscan is the best documented language among Paleoeuropean languages, a group of at least 3 certain members (Etruscan, Rätian, Lemnian), but maybe as many as 7 members (Etruscan, Rätian, Lemnian, Elymian, North Pikene, Lepontic, Camunian). Central Paleoeuropean is tentatively assigned to the central group, because the word *mala* meaning "still water, reflection" is found in it, this word is the basis for the Etruscan word *malstria* for "mirror". If the three areal language groups should be considered to be derived from the same stock, then the timedepth for unity would be in the range from 10,000 to 20,000 years before present. The areal pattern suggests that the unity of the Paleoeuropean languages were broken up by the advance of early Indoeuropean languages, probably in the time-period from 7,000 to 4,000 BC.

## 8 Conclusions

We should be careful with our conclusions at this point, but we would like to make the following points:

\* Of the non-Indoeuropean remnant languages called Paleoeuropean, sufficient material for a certain genetic classification is only available for Etruscan, Rätian,

Lemnian and Iberian. For all the other languages listed, the material is too limited and fragmentary

\* A comparison of the extinct Paleoeuropean languages show many similarities in the non-cultural lexicon. Some of the languages show similarities in numerals and pronomina.

\* The areal pattern suggests that the hypothetical unity of the Paleoeuropean languages was broken up by the advance of early Indoeuropean languages, probably in the timeperiod 7,000 to 4,000 BC.

CAMUNIAN			
Word	Translation/comments		
		χezpaz	?
		χusus	Etr. <i>cusus</i> "drink"
		leima	female name ?
		leimi-ez	female name ?
alai-al-z	name	lath-i-al	name, Etr. <i>-al</i>
a*renz	?	lanether	?
i-asaz-iz	Etr. <i>ais</i> , "God"	minaka-i	name
aststaz	?	munth-al	"infernal", Etr. <i>munthial</i>
eltnaz	?	nio*is	?
eko	IE, <i>ekom</i> , "I"	neak-en	?
enotinaz	?	nenz	?
esai-e-al	Etr. <i>-e-al</i>	prianz	?
esi-al	Etr./Rät. <i>-e-si-al</i>	pueiaz	?
ezt	?	pinz	?
iunke	?	pinz-t	Etr. <i>-t</i>
itz	?	pinz-ti	Etr. <i>-thi</i>
iueica	?	le-pap-al-z	Etr. <i>papalser</i> ,
χem	PaleoE, <i>kem</i> , "lake"		"grandchild"
χem-al-az	name		

le-pali-al	Rät. <i>pali</i> , Lep. <i>pala</i> , "grave"	χis	Bsq. <i>gison</i> , Ib. <i>gisen</i> , "man"
le-	prefix <i>le-</i> , noun class ?	-e-mi	pronoun, "am I"
φrziukitulkφ		-en	possessive suffix ?
rethe-n-al	Etr. <i>-al</i>	elymoi	ethnonym, Elymians
supre	verb ? Rät. <i>surpri-khe</i>	ery-k-a	toponym, "the city", Bsq. <i>iruna</i>
tine	? Etr. <i>tine</i> , Rät. <i>tine</i> , "lead"	eru-k-a-zi-e-mi	"of the city itself am I"
u	?	ila-ai	?
uelal-al-z	?	iru-k-a	Toponym on coin, "the city"
uekez-us	name ?	iru-ka-zi-ie	toponym, "of the city"
umuin	?	-k-	ergative ?
uelai	?	-ka	locative suffix, Bsq. <i>-ko</i>
uati-az	?	kim	PaleoE <i>kem</i> , Etr. <i>cem</i>
ui-al-k-im	?	kuti-ai-e-mi	?
ui-pi-al	?	-la	Cauc/Etr. arch. <i>-la</i> , rez. <i>-al</i>
zepkas	? Etr. <i>cezp</i> , <i>semph</i>	lena-ai	?
zezak-ai-z	?	lep-an-a	?
zazi-al	? Etr. <i>-al</i>	logash...	?
ELYMIAN		lutne-la	Etr. <i>lautne</i> "of the family"
Word	Translation/comment	-mi	possessive, "mine" ?
at-ai	? Etr. <i>ati</i> "mother", Etr. <i>athe</i>	mlak	Etr. <i>mlakas</i> "beautiful"
aths-en-a	? Etr. <i>athe</i> , "house"	muta-ai	Etr. <i>muthina</i> "netherworld"
atroi	name ?	nata-ai	?
		pid-ai	?

puke	?	acn-	produce, create, make
pinas	?	acnanas	foster, raise children
panna-ai		acnasvers	foster the (sacred) fire
sariχ	Etr. <i>sacer</i> , Ib. <i>sakar</i> , "sacred"	acil	thing, necessity
sariχ-uakar	Bsq/Iber. <i>urke</i> , "height"	+acale	month of June
silia	?	acvil	gift
segesta-zi-e-mi	Coin legend, "Segesta from am I"	ais, eis	God
thasid-ai	?	aiser, eiser	Gods (plur)
thi-ai	?	aisiu	divine, with the gods
titel-ui	demonstrative, "similar to"	ais-na	divine, of the gods
titel	demonstrative, "of this, name"	aiseras	of the Gods
titite-la	name, "of titite"	al, alie	verb, give
tulai	Etr. <i>tular</i> , Rät. <i>tulie</i> "border"	aliche	passive, verb, was given
tuk-ai-e-mi	?	alike	active, verb, was given
tusk-ai-e-mi	?	aliqu	verb, would have been given
-u-akar	Etr. <i>ukar</i> , Ib. <i>urke</i> , "high"	alias	noun-genitive, "of the gift" ; symbol of the real object
zelile	Etr. <i>zilath</i> , <i>zichl</i> ?	alpan	to please, make glad
		alp-nu	gladly
		alph-as-e	happiness, pleasure
		als	sea (salty) Grk. <i>halos</i>
		als-ie	sea port, Lat. <i>alsium</i>
		als-as-e	at the seas
		alumn	society
		alumn-athe	society house (sacred ?)
		am, amke	verb; to be, was

# ETRUSCAN

## Word Meaning and comment

### A

-a	dative singular suffix
-as	dative-genitive suffix
acas, acazer	make offer, offers

amucu	verb; will be as, conjunctive	ati nacna	mother-fosterer; grandmother
amuce	verb; has been	avil	year
amþe, hamþe	month of May	avilcva	annually
an, ana, ane	pronoun, he, she	avilcval	of the annual
anc, anac	pronoun, he, she	avilsch	annual
anai, anaies	relative?, pronoun; his? Bsq., <i>anai</i> "man's brother"	a-nunthe-ke	the made sacrifice
		a-nethe-ke	the gift given
*ande, *unde	large, Bsq. <i>andi</i>		
+antha	eagle, north wind	C	
apa	father	-c	and
apana	on behalf of the father, fatherly	ca	this
apasi	of the father	ca-l-thi	this of here
apcar	abacus	camthi	title; dictator
afrs, apries	april ?	canzate	adjective ?
apir-as-e	spring ?	cape	vase
ar, arce	make, build, do	capi	verb; take away
aras	flight	+capr	month of April
arasase	air	capra	urn
+arac	falcon	+capu	hawk
+arima	the monkey	*cara	stone
ars	push away ? remove ?	casti-al-th	
-as-e	amount-word, "-stuff"	cautha, catha	sun-god
aska	Greek vase	cer-i	verb; build
athe, athre, etre	house	cer-i-chun-ke	was built
ati	mother	cer-i-akve	erect votive, "build gift"
ativu	motherly	cehen	this, it, here
		cekha	verb; to want, will

cekhana	council, assembly, board	cve, cve-r, kvil, cve-r-a gift
cela	room	chun, cun-ce verb; set, was set
+celi	month of September	+chospher month of October
celu	priestly title	*cezpze 7th
cep, cepen	priestly title	cusus ritual drink
cep-ta	priestly ?	quthefas revenge, victory
ces-, cesu	verb; lie, rest	E
cezp	7	eca, ecas, ecal this, of this, from this
cezp-alch	70	ecnia every
ci	3	ecisie amphora
cialch	30	eithva, etva big, large ?
cialath	for the third	ein no, not
ciz	thrice	eleivana olive oil
ciem zathrumis	17	elssi then ?
kisnee	third	-em less, taken from
kihak	stone	enac then, afterwards
cilth	castle, fort ?	eni as
chimt, chimtm	hekatombe	eniaca as these
cirra	man, youth	epl (e-pul) in, to, up to
clan, clenar	son, sons	er, erke make, erect
cletram	basket	+erminius month of July
cleva	offering, sanctuary	ez, eslz (e-zal-s) twice
cluvenias	sacred precinct ?	ezlem zatrumis 18
creal	magistrate	ethvi partner in marriage
cule, culs-cva	door, gates, ?	+etera, eteri foreigner, slave
culichna	kylix	etnam and, also
cupe	cup	eth in the presence
kurpi	verb; make, wrought	ethr-, etr- religious action?, house?

-eska	hand, power	fulumcva	all the stars, heavens
eskat	verb; take (with hand)		
eskatke	was taken	Φ, Ph	
evi-tiuras	thus of the monthly	+phersu	mask, face
		phersun	person
F		ph---	1000
fala	look, see	phur	1000 ?
falathi, falanthi	heaven	phuris	thousand times ?
+faladu	sky	phechucu	portray ?
falza	lake		
falzathi, falsti	at the lake ?	H	
fan	fire, consecrate	hamøe, ampile	month of May
fanu	sacred place, on a hill	hanthin	in front of
farth, far-i-ce-ka	bring, brought	hece	put, add, arrange (verb)
farthne	bringer (girl), Grk. <i>parthenos</i>	herma, hermasva	statue, statues
favi	grave, crypt	+hermi	month of August
fas	because	hil, hilar	youngster, man ?
fasti	because of	hilar nesl	the dead youngsters
fasle	vase	hinh-a	below, netherworld
fase	sacrifice	hinh-u	infernal
fler	sacred picture or statue; name	hinh-i-al	of the netherworld, soul
fler-cva	portraits, votives	+hiuls	owl
flere	the divine appearance, divinity	hupni	cemetery
front	thunder	hupnina	belong to cemetery
front-na-c	diviner of thunder	hus, hus-i-ur	child, children
fulum, pulum	star	husn-atre	children-house, -- of the house
		husina	youth
		huth	4

huvi	noun	lauchum-na	of the king; palace
huvi-tu	noun	lauchum-ne-thi	royal month of November
I		lachuth-, lacth	royal offering
ik, ikh, ic	as	lautun, <sup>+</sup> lautni	family
ikhnac	how of	lein	verb; die ?
ikl	as of	lekhtum	vase
ika	this	<sup>+</sup> leu	lion
icac	as these	les	silver
ik-u-tevr	...and as the bull...	lucair	rule
ilu, ilace	to pray, prayed, favored	luth	sacred place
ilacve	prayers, favors, active	lup-	verb; to die
ilachve	prayers, favors, passive	lupnu	dead
iku	a prayer	lursth	offering after the census
in, ink	it	laur	land
ipa, ipe	whoever, whatever	laur-s-ti	from the land
ipi	here, in	laursti-al	from the land's
ipul	in, upto (inanimate)		
<sup>+</sup> ister	play actor	M	
*is-	river	-m, -um	but, and for, inanimate and
ita	this		
*itu	divide ?	mach, mak	5
ittun	container	machstrev	magister, learned person
L		mal	show, picture
las	offering	mala	reflection, image
las-te	stone plate, "offer there"	malena	mirror
		malstria	mirror
lauchum	king	mala-vis-ch	makeup, show adorned



man, mani	the dead, Lat. <i>manus</i>	munth	world of the dead, tomb
manijim	but for the dead, monument	munthial	in the tomb
mantrans, man-tran-s	memorial speech	munthna	funerary
marni	title, profession ?	munistas	causalties, deceased
maru, marucva	quaestor, lawyer	mur	stay, reside
masan	muse, divine	murs	urn
matnam	above	muvalch	50
mathu	mead, honey	musa, mese	divinity
mathcva	drunk, full of drink	mutana, muthna	sarcophagus
mech, mechlum	people	menitla	adjective
*meit, methlum	community	mnehvra	amphora
men, menake	verb; offered	N	
mese	muse, divinity	-ne	personal pronoun
mi	I, 1. pers. pron.	nac	how
mini	me, 1. pers. pron.	naces	of how
mim-enica-c	and this memorial	naplan	wine jar
mlak, malake, mulach	beautiful, good	nefts	nephew
mlakuch	would have been beautiful	nene	nanny
mlakuta, mlachta	beautifully	neri	water
mlerusi	brave ? good compatriot	a-netheke	noun; type of offering
mlati, mlatke	beautify, make good	netheke	verb; ? was offered
mul	gift, present	nes-na	destroyed ?
mulu	verb; dedicate	nesl, nes-al	of the dead
mulveni	dedication	nes	dead, deceased
mul-s-le-mlach	the good gift	netvis, nethsra	divination of lightning
mun, muni	world	nuna	the offering
		nuntheke	verb; offered
		nurph	9

nurphzi	9 times	R	
nurphalch	90	ras	noun; human, man
nuz	sheep	ras-na-l	of Etruscan's, "of the men"
P		rac	noun; bird, falcon
papa, papa-c-s	grandfather	rakt	prepare (bird sacrifice)
papa-l-s	grandchildren	rasenna	Etruscan ethnonym, "belong to the manly"
parnich	magistrate	rath, ratac	law, sacred tablet
patna	vessel, vase	ratum	according to law
paites	at home, in the house	ril	age
penthuna	stone slab	ruva	brother
peras	rich		
pethereni	month of December	S	
pi, pul	in, through	sa	6
pop	knee	sam	6th
prumats	greatgrandson	-sa	"son of " with a name
prukum	vase (from Grk.)	sal	2
puia	wife	sac	consecrate
pulum, pulum-kva	star, heavenly	sacni	sacred
palan-thi, falan-thi	sky, heavens, stars place	sal	verb; carry (water) ?
purth, puru	head, Pict <i>brude</i> , Bsq. <i>buru</i>	+sant	metal
purthenna, porsenna	leader, (cf. Lars Posenna)	sanisva	metallic ?
puruhena	portrait ?	sar	10
put	well, fountain	scune	dispose
putes	of the well	scunsi	of the inheritance
		sealch	60
		sec, sech	daughter
		sela	verb; shine

selace	verb; shone	T	
seleitala	"this the glory of"	ta	this
semph	8	tamera	title; judge at court
semphalch	80	tamer-eska	obligation, memorial
sliaches	society		grove
sneath	maid, companion	tham	verb; to found
snuiaph	in multitude	thamuce	verb; will be founded
spanti	(stone) plates, altarstone	+tamna	horse
spet	drink	tanasa, thandsa	actor, dancer
spur	city	than-cvil	thana-gift, beautiful
spurana, spurieni	urban, sing, plur.		woman
spuriaze	public	thapna	vase
spural	townsman, citizen	thar	100
sren, sran	stature, ornament	thaurc	funerary
srencva	ornamented	thaur, thaura	tomb
srencve	ornaments	tarils	deity name ?,
suc	verb; declare	tarch-na-l-thi	Tarquin-of-from-there
sucri	to be declared	thafna	cup (in general)
suplu	flutist, suppliant	temia, tmia, tmial	temple
suth	stay, rest	themi-asa	religion
suthi	resting place, grave	ten-, tentas	to act as, lead (verb)
sval, svaltas, svalce	verb; lived	tera, teras	animal, animal's
sve	similar	te-, tei	to care
surieisteis	Suris, eis=God, teis=care of, "In care of the God Suris"	tes-, teis, tesam	to care for (verb)
		thesia	caretaker, bureaucrat
		thesia-meit-al-e	on behalf of the popular administration
		tesinth	care-house; office

tesnsteis	tes(a)n-(ei)s-teis, Tesan-god-care-of	tularis	of the borders
thez, thezie	to make offering (slaughter)	tunur	one at a time
thez-ri-c	make offering-should-and	thu, thun	1
teta	grandmother	thuva	single
tev	to show, set (verb)	thunz	once
tevarath	member of audience herald for lituus	thunur	single
tevru, thevru	bull (Semitic)	thunem zatrumis	19
thi, thir	we, they, 1. pers. plur. ?	Turan	Goddess Venus
tin-	Jupiter, day	turu	give, Bsqr. <i>i-torri</i> , well
thina	earthenware pot	turike, turuce	given, would have given
tiu, tiv, tiur	month, moon	tura	gift-the, incense
tiur-uni-as	the month of Uni's	turza	offering
tis	east, sunrise	tusu	pillow, resting place
trane	month of July ?	tusur-thir	on pillows there, married couple
tras, trasce	to become, became (verb)	tuthi	state, tribe
-tras	member of ---	tuthina	citizen, of the tribe
-tur	member of ---	tuthiu	tribal
trin	plead, promise (verb)	thucte	month of August
trinake	spoken Rät. <i>trinake</i>	thui	here
trouna, thruna	rule, govern	thuni	here, now
trut, truth	libation, spell	tva	verb; show
trut-nut	priest, oracle	U	
tul, tular	wooden post, gate, border	-u	verbalsubstantive suffix
tuler-ase	all the borders	u-ne	then
		u-sil	noun; sun
		u-s(il)-la-ne	at noon

ut	perform	zichl	the book's, text
ukar	high, fortress <i>u-kar</i>	zichu	scribe
	"high stone"	zil	administrate, lead
		zila, zilc	magistrate, president
V		zilathi	presidency
vacal, vacl, vacil	libation, oath	zina	give (verb)
vati	to arrange, accomplish	zinake	have given
vatie-che	have arranged, have	ziv	the "have lived"
	taken on	ziva, ziva-i	the dead
+velcitna	month of March	zina, zinake	given
vertun	vase	zusle, zusleva	victim, animal
vers, fers, phers	fire	zuci	declaration
vel	knight, sir		
vinum	wine	Numeral	n
		thuva	single
Z		ez, eslz, e-zal-s	twice
zal, e-sal	2	cis	three times
zelur	doubles	huthis	four times
zathrum	20	machs	five times
zathrum-s-ne	20th	saris	ten times
zatlath	companion	phuris	1000 times
zar	10		
zaris	10th	cizi	3 times
zavena	drinking vessel	nurphzi	9 times
zeri	something related to		
	write, letter ?	thunur	singles
zich	write ! passive	zelur	doubles
zik	write ! active	celur	triples
zicuke, zichuche	written, was written		

thun-s-na	of the first	cis cealchls	33
ki-s-ne-e	of the third	huths cealchls	34
cezp-re	of the seventh	huthachl	40
zathrum-s-ne	of the twentieth	thunem muvalchls	49
		muvalch	50
Numeral n		cis muvalchls	53
thu	1	huths muvalchls	54
zal, e-sal	2	sealch	60
ci	3	machs sealchs	65
huth	4	cezpach	70
mach, mak	5	esals cezpachls	72
sa	6	semphalch	80
cezp	7	machs semphalchs	85
semph	8	nurphalch	90
nurph	9	tha	100
sar	10	ph(ur)	1000
?	11		
?	12		
ciszaris	13		
huthsaris	14		
ciem zathrumis	17		
ezlem zatrumis	18		
thunem zatrumis	19		
zathrum	20		
huthis zathrumis	24		
ciem cealchls	27		
ezlem cealchls	28		
thunem cealchls	29		
cialch	30		

	1.pers neutral	1.pers class e-	3.pers neutral	3.pers animate	3.pers inanimate
Nominative	mi	e-mi	*na	a-na	i-na
Absolutive	mini	*e-ni	.	a-na-i	*i-na-i
Ergative ?	.	.	.	*a-na-k	*i-na-c
Nom. ? plur.	ti	*e-ti	.	.	.

Table 1: *Observed and reconstructed archaic Etruscan system of personal nouns.*

cases	stem	Class e-	Class i-	stem	Class e-	Class i-
Nominative	ta	eta	.	ca	eca	ica
Accussative	tan	etan	itan	cn	ecn	.
Genitive I	ts	.	itas	cs, cas	ecs	.
Genitive II	tla	.	itala	cla	.	.
Ablative	?	.	?	cs, ces	.	.
Ablative II	?	.	?	clz	.	.
Pertinentive	?	.	itale	cle	.	.
Lokative	?	.	italte	clt, clthi	ecclthi	.
Pertinentive II	.	.	.	.	ecnia	.

Table 2: *Observed Etruscan system of demonstrative pronouns "this".*

Meaning	Stem	Class e-	Class a-	Class i-	Class u-
and	-m	e-m	a-m	i-m	u-m
2	zal	e-zal	.	.	.
1	thu	e-thu	.	.	.
how	nac	e-nac, e-nac-es	.	.	.
sun	sil	.	.	.	u-sil
offering	nunthece	.	a-nunthece	.	.
man, official	lace	.	.	i-lache	.
?	chu	.	a-chu	i-chu	u-chu
in	pul	e-pul	.	i-pul	.
into	pi	e-pi	.	i-pi	.
wife	puia	.	.	a-puia	u-puia
leader	purthne	e-purthne	.	.	.

Table 3: *Examples of noun prefixes in Etruscan. A preliminary guess at the function would be e: animate personal, a: animate, i: inanimate, u: collective, inanimate.*

IBERIAN		anaio	brother ?
		andi	large ? woman ?
A		ar	verb; to be
-a	determined article	ar-mi	1st pers. I am
-ai	demonstrative	ar-ban	verb, impersonal; one is
agin-irte	kill	ar-ku	verb; thou art
abarr	hand	ar-nai	signature "am I"
abar-tan	hand-from-the; dance	arre	demonstrative; here,
ade, ate, ati	father		pronoun; "man" or
adin, atin	elder		verb; "lies" ?
ake	mother ?	argi	light ?



arse	fortress (Sagunto)	bask	noun, food ?
as	god, spirit	be-laka-s-ik-aur	by the smaller
asgandis	"of the great spirit" ?		magistrate ??
-aur	relative, child of ?	be-laur	little field ?
aur-	soil, mud ?	beles	black
		berr-en	of the new (genitive)
B		berr-i	new (dative)
ba	1	bi	2
-ba	the (-one)	bide	path, track
-ban	the one, det. art, "son"	bid-en	of the road
-ba-n-en	of (the one) poss.	bikir-laku	magistrate title
-banite	..of the ones..	bin	together, unite ?
ban-	demonstrative article,	bin-ike	united-(ergative)
	one	biur	twisted
ba-garo-k	the harvest ?	biur-lakos	attribute-title
bai-d-es	in the house, at home	biskar	hill, ridge
bai-d-es-ir	in the community house	bisku-te	mined, extracted from
bai-d-es-gi	in the house there		hill
bai-d-es-ban-i	in that house	bilos	good
balke	community, people	bod	lake, pond
balk-e-laku	community magistrate	-bor	head, leader
balk-e-lako-sk-a	popular magistrates	bor-ste	much; hand-full ?
balk-ar-isk-ar	chieftain	buistin	sandy
balke-biur-ai-es	people twisted ? ?	buko	cape, promotory
bar, ibar	valley		
baser	sacrifice	D	
basirtir	at the cementary ?	dun	darkness, holy ?
	at the settlement ?	dadula	deliver ?
bas-i-balk-ar	on behalf of the people?	u-duin	sufficient ?

E		erre-nai	verb-pronoun ?
e-ba	one	erre-su	verb-pronoun ?
e-ba-n	one the	erre-ta	verb-pronoun ?
e-ba-n-en	one the of; child, son	erre-te	verb-pronoun ?
e-barr-ik-am-e	with my own hands ?	erter	verb
ede-	take away	e-, es	house ?
ede-silir	take away silver	euki	noun, possession
ede-tur	travel across border, get water ?	eukiar	verb, possess, own
ede-sike	name ?	ezker, isker	hand, powerful
ekou, eku	produce	G	
eguan, ekuan	produce, product	-k-	suffix, ergative ?
ekarr, ekorr	verbal-substantive, product	-ki	with
eki, ekiar, t-ekiar	verb, make	garo	cereal, grain
ekiar-done	verb, make darkness ?	garo-k	grain-plural
eko-te	verb-pronoun; they made	garo-k-an	of the grain-plural
er	verb, existential meaning	-gi-ba-s	from-one-of ?
erke	verb, build ?	gis	man
erker	verb ?	gis-en	of the man
er-ir-il	verbal form ?	gudu	battle, strife ?
er-ir-tan	verb; lived ?, aged ?	gudu-boike	fallen in battle ?
erre	verb-?	I	
erre-is	verb-pronoun ?	-i-	infix
erre-ko	verb-pronoun ?	ibai	river, stream
erre-ku	verb-pronoun ?	i-ber, e-ber	much, rich
		iberar	"iberians, "the rich ones" ?
		ikon	noun; fear ?

ikor	name (verb), panic ?	K	
ildu, ildun, ildur	(death)-darkness	-k, -g	plural suffix
-ike, -ig	ergative suffix ?	karr, karri	stone
ila	death, sacrifice ?	-kerr	stony
il, ili	city	katan	the path, the road ?
il-i-berr-is	city name, "the new city"	kelde	forest ?
		keltar	wooden, forest ?
ili-(i)tur-ki	placename, city-well-at	kem	swamp, wetland
iltir, iltir-ite	city, cities	kerr	build in stone
ir, ire	community	ki, kiao, kior	3?
ir-a-en-ai	the citizens of...	-ki	derived from location
irt	death ?	-ki-bas	-from-one-of
-ir	pronominal suffix	-ko	belonging to location
-irr	flectional suffix	-ko-ba	one belonging to location
is, isar	water, river		
isai	river	ki-ta-r-un	300 ?
isker	hand, power	kidei	decide ?
isbatar	name, "the one from the river" ?	kokor	hard stone, cliff, rock, cruel ?
-ite, -te	plural suffix	koroti	verb, accomplish
itur	well, spring	koroite	verb, accomplish, plural
itur-ir	placename, "well-town"	goroti-gi	verb, accomplished
iu, iun-	pronoun ? our ?	kotor	cliff, mountain
iunstir	our city ?/lord of - ?	ku	pronoun; you
iunstir-laku	title, leader of IUNSTIR	-ku	pronominal suffix, 2. pers. sing.
i-agin-ur-e	verb; cry, it-make-water-him ?	ku-s	demonstrative-pronoun; from it
		kares-ban-ite	name-one-(plur)

## L

lako, laku, lake	magistrate
laku-sk-i	leaders (pl.)
laku-iltir	city magistrate
laku-ar-gis	by the leading man
lau, laur	field land
legus-egi-ka	made by junior magistrate ?
lez	metal, lead, ore
lu	4 ?

## M

mbarr	gift, "??-hand", 10 ?
m...	5
mek	5 ?
mi	I, mine
mkei	adjective, good ?
n-mkei	noun ?, the good ?
mliki	good, beautiful ?
mlirr	?
mun	land, area

## N

nabar	extraordinary
nal-	name prefix
nai	I, me
neitin, neto, neton	name
nere	myself ?
neska, neso	girl, daughter ?

## O

oasai	name
oisor	scream, cry ?
okar	goat ?
otar	building
oto,	verb, build ?
otoke	was built

## S

sabari	placename, cattle pen ?
sakar	strong, sacred
sakar-lako	title, strong leader ?
sal	river, stream
saldu	noun, market, sale, trade
saldu-ban-ite	customers, market ones
saldu-laku-ki-arr	leader of salesmen there
salir	silver, value, money
salir-k	payment
san-i	verb, thank
san-ir	verbal conjugation
san-er	verbal conjugation
san-u-ke	verbal conjugation
san-i-ke-ai	verbal conjugation
san-i-ke-ar	verbal conjugation
san-i-bar	verbal conjugation, thankful ?
se	10

seltar	grave	t-eban	his/her son, his/her
selke	was written ?		child
selkisker	scribe	t-ebanen, t-e-ba-n-en	his/her child's
semr	kinsman ?	-tegi, -teki	toponym suffix, "here"
seni	son	-teko	toponym suffix, "here"
sere-meki	50th ?	-ti	derivative locative
sike	verb, fill ?		suffix
sorse	numeral ?; 8	-tiba	there-one
sosin	just, truth	-ti-e-ba	there-it-one
sosin-biuru	lies; truth twisted	-ti-e-ba-n	there-it-one-the
starien	of the rich	tikir	prince, Celt <i>tigirnos</i>
ste	amount	tolir	Etr. <i>tular</i> ; borders
-sk-	plural suffix	tur, itur	spring, Bsq. <i>itorri</i>
-sk-a	plural determined form	turl	mountain pass
-st	plural, abundant		
-st-a	plural determined form	U	
-su	pronoun (you, plur.)	udaor	harvested, extracted
-se	pronoun (he, they)	ur	water
suniar	noun	urke	high, fortress
		ustai	Bsq. <i>ustai</i> ring
		-un	noun suffix
T			
ta	numeral 100		
t-a-ke	verb, he-is-here	Numerals	Comment
t-e-ke, teke	he-is-here	ba	1
t-e-ike-o-en	he-is-ERG-?-of	bi	2
t-a-gis-garo-k	he-is-man-cereal-ERG ?	ki, kiao, kior	3 ?
-tan	related to -tar ?	lu ?	4 ?
-tar	geogr. derivation suffix	mek	5
		se	10

abarr-ike 11 ?

ta 100

## LEMNO-PELASGIAN

aker Etr; ukar high

arai verb ? he made ?

aomai noun ?

avis Etr; avils years

eptesio adjective

evistho ?

fala see, look

foke name Foke

foki-as-i-al-e name-GEN-DAT-GEN-DAT the Foke's's

haralio adjective ?

holaies name Holai

holai-es-i name-GEN-DAT of Holai's

hytt numeral, toponym, Etr. *huth* 4

maras Etr; machs, Ib; meki 5

maras sialkvis Etr; *machs seachls* 65

marasm sialkveis 65

marasm 5th

mav Etr; muv-ach-l 50

morinai-l name-GEN Morina's

nafroth Etr; nefts nephew

seronai placename on Lemnos  
placename

seronai at Serona ?, prefecture ?

sialkvis Etr; seachls 60

sialkveis 60

siasi name Si

sivai Etr. ziva, zivai the dead

tis Etr; tis (give, lead, determine)  
give ?

tavarsio adjective, Etr; tev  
appearance ?

toveronai Etr; tevarath audience,  
electorate

thamesa Etr; thamereska obligation

vanal-s-i-al Greek; *wanax* of the king's

uar Tart; uar ?

## LEPONTIC

alias IE. suffix -os

alkoun-os name, IE. suffix -os

askone-ti name

amasi name ?

asmin-a female name, IE. Suffix  
-a

dieu ? Etr. *tiu*, "moon"

eis-na divine, Etr. *aisna*

kasil-os name, IE. suffix -os

kois-ai name

komone-os name, IE. suffix -os

krasan-ik-na female name

kual-ui name

latumar-ui	Latu the official,	NORTH PIKENE	
metel-ik-na	female name	ait-en	?
metel-ui	name	arnuis	?
maesil-al-ui	name	anos	her ? Etr. <i>an</i> he, she
min-u-ku	name, pronominal form?	bales-t-en-ak	?
min-u-i	name, pronominal form?	et	demonstrative ?
nas-om	?	eus	?
pala	"hollow, grave"	erut	?
pe	Etr. relative pron <i>ipe</i>	isper-i-on	?
pevak	"enactment"	isa-i-on	Ib. <i>is</i> , water
piuo-ti-al-ui		ipi-em	Etr. <i>ipi</i> , in
piuo-n-ei	?	kaarestad-es	?
pruks	?	kalatne	name
ranen-i	name	krus-t-en-ak	?
sap	pronoun, 2. person ?	laku-t	Etr. <i>laucumo</i> , Rät. <i>laku</i> , king
slani-ai	name	lutu-is	?; Etr. <i>luth</i> , sacred place
sillok-ui	name	mim	memorial? Etr. <i>maniim</i>
sutai	grave	merpon	mer-p-on, 5th ?
teki-al-ui	?	nesi	Etr. <i>nes-l</i> , of the dead
ter-om-ui	name	nis	Etr <i>nus</i> sheep, <i>nes</i> , dead, Bsq <i>ni</i> , I
tisi-ui	name	pat-en	Ib <i>ban</i> , Bsq <i>bat</i> , Pict <i>pant</i> , I
trau	?	part-en	?
vala-un-al	name	pol-em	?
veni-a	family ?, son ?, wife ?	rot-em	Etr. <i>ratum</i> according to law
verk-al-ai	?		
vin-om	wine		
uletu	?		
uasam-os	name, IE. suffix <i>-os</i>		
util-os	name, IE. suffix <i>-os</i>		

rot-n-em	?	<sup>s</sup> karr	"stone", Etr. <i>ceri</i> , Rät.
rot-n-es	of the law ?		<i>karr</i>
sut	Etr. <i>suthi</i> , grave	mezunemusius	name, IE. suffix -us ?
sot-er	graves ?	<sup>s</sup> pali	"hollow, grave", Rät.
sot-r-is	of the graves ?		<i>pali</i>
tet	?	<sup>s</sup> pod	"lake, tarn", PaleoE. <i>bod</i>
thalu	?	<sup>s</sup> zupar	"hide, conceal" Bsq
trat	?		<i>zopar</i>
trut	Etr. <i>truth</i> libation, Celt.	<sup>s</sup> susin	"fair, even" Bsq. <i>zuzen</i>
	name <i>Drut</i>	<sup>s</sup> taur	"mountain"
teu	Etr. <i>tev</i> show, <i>tiu</i>	<sup>s</sup> tziagarru	"dog", Bsq. <i>tzakur</i>
	month, moon	uvezaruapus	name, IE. suffix -us ?
tasur	?	vemetuvis	
tisu	?		
tret-en	?	NURAGIC	
teletau-n-em	?	<sup>s</sup> baites	"in the house"
vult-es	name ?	es	house
us	?	is	"water", Bsq. <i>is</i>
		<sup>s</sup> iska	"river"
LIGURIAN		<sup>s</sup> istula	<i>is-tula</i> ; "riverbank"
akiu		<sup>s</sup> karr	"stone"
at	"house, dwelling" Etr.	<sup>s</sup> kita, kide	Bsq. <i>kide</i> , comrade
	<i>athre</i>	-ni	pronoun ?, Bsq. - <i>ne</i> , - <i>ni</i>
a-rus	"man" ? Etr. <i>ras</i> , Rät	nurgo	?
	<i>arus</i>	sardo	Ethnonym
<sup>s</sup> baita	"in the house", Rät.	<sup>s</sup> tzakurru	"dog", Bsq. <i>tzakur</i>
	<i>baithe</i>	tula	"side", Etr. <i>tular</i> ,
<sup>s</sup> bera	"rich" Rät. <i>bera</i>		"borders"



# ITTURIAN

<sup>s</sup> ali-st-os	"wet forest land, alder"
<sup>s</sup> -andr-	"large" ? Bsq <i>handia</i>
<sup>s</sup> balv	"hollow, cave"
<sup>s</sup> is-	"water, river", Bsq. <i>is-</i>
<sup>s</sup> itter	"spring", Bsq. <i>iturri</i>
<sup>s</sup> sal	"stream, river"

# NORTHWEST PALEOEUROPEAN

<sup>s</sup> -a	definite article suffix
<sup>s</sup> -andr-	"large" ? Bsq <i>handia</i>
<sup>s</sup> -ak	suffix
<sup>s</sup> -am	suffix
<sup>s</sup> -an	suffix
<sup>s</sup> ald-, -alde	"side, area", Bsq. <i>halde</i>
<sup>s</sup> ali-st-os	wet forest land, alder
<sup>s</sup> balv	hollow, cave
<sup>s</sup> bal-sa(l)	puddle, Bsq <i>baltsa</i> "pond"
<sup>s</sup> bara	"valley", Bsq. <i>ibar, bar</i>
<sup>s</sup> bantia	1, the one, Pict. <i>bannatia</i>
<sup>s</sup> bod, bod-am-a	"lake, at the lake"
<sup>s</sup> borm	"well, spring"
<sup>s</sup> bisk	"height", Bsq. <i>*bisk-al</i> ridge
<sup>s</sup> bris	?
<sup>s</sup> -ink-	suffix
<sup>s</sup> is-	water, Bsq. <i>is-</i> ; "water"

<sup>s</sup> itter	"spring" Bsq. <i>iturri</i> ; "spring"
<sup>s</sup> kallan	forest, Pict <i>kaled</i>
<sup>s</sup> karra	stone, Bsq. <i>harri</i> , Etr. <i>ceri</i>
<sup>s</sup> kisim, kiem, kem	"lake, waterbody"
<sup>s</sup> mun, mund-	land, Etr. <i>munth</i> , "world", Bsq <i>muna</i> , height, land
<sup>s</sup> meth, med-	settlement, Etr. <i>methlum</i>
<sup>s</sup> -nth	suffix
<sup>s</sup> -pa	the one ? suffix
<sup>s</sup> sal	stream, Bsq. <i>sal</i> , "water"
<sup>s</sup> -st	suffix
<sup>s</sup> tsom	salt
<sup>s</sup> ur	water, Bsq. <i>ur</i>
<sup>s</sup> var	river
<sup>s</sup> verm	large river

# CENTRAL PALEOEUROPEAN

<sup>s</sup> -a	definite article
<sup>s</sup> -ander-,	large Bsq. <i>andi</i> , Etr. <i>ande</i>
<sup>s</sup> karra	stone, Bsq. <i>harri</i> ; stone
<sup>s</sup> katar, kotor	rock
<sup>s</sup> mala	calm water, reflection
<sup>s</sup> -nth	name suffix

PICTISH		<sup>+</sup> clar, <sup>s</sup> clan	children
A		<sup>+</sup> cartit	brooch
all	hill		
allhhalldorr	hill of Caledonia	D	
atadd-o-ar-en	name, Atadd's	dunnodnnat	name
am	he is	dattrann	norse loan, daughter's
n-ahht	the mother	ddroan	name Troan
		drosten	name Tristan
B		<sup>+</sup> diu-per	rich
<sup>s</sup> bant	numeral 1	<sup>s</sup> deitu	scream
<sup>s</sup> bantia	placename market place	<sup>s</sup> diu, <sup>s</sup> deo	moon
bernises	name, Bernice's	E	
belancen	name, Belan's	edd, ett	dead
bedra-c-l-o-v-or	name, Bedra	ett-or	the dead
<sup>+</sup> buthib	slave	etettfor	name, Etett
<sup>+</sup> brude	leader, head	ems	verb ? ?
*blies	adjective. black	n-ehht	the father
<sup>+</sup> but, *bet	small, younger	in-ehht-es	her father
		ah-ehht-an-n	his father's
C		eoðto	name, Odd
cerr	noun stone	ectuc	
<sup>s</sup> cairn	stone heap	ehtuc	by the father ?
cerroccs	cross	ettecon	of the dead ?
crrosc	cross	eddamon	title/name Ethernan
cus	cross	ettocuhetts	name, Ettocuhett
u-cota	knife	o-evv	stone
caddosten	name, Caddost's	n-evv	stone
caled	PaleoE. <i>kallen</i> forest	e-evv	stone

# COMPILING WORDS FROM EXTINCT NON-INDOEUROPEAN LANGUAGES IN EUROPE

hccvv-evv	stone	movvest	name Mofest
		maqq	son of
G		maqq-o	son of
+guern	good	m'qq	in the family of
+gungoch	men	meqq	in the family of
gariost	name Gariost	meqq-n-an	of the family of
gis	man	meic	belonging to the family of
H		+*mioch	mine
hhalld, caled	forest	+morduit	name, Mordred
+hib	*kib boy, man	mats---	name
		morsa----	
I		mund	land, soil
ipi, ipe	sister-son	mundnaith	name, "from the land"
ipu-or	the sister-son		
ipe-u-or	the sister-son	N	
ipe-nn	the sister-son's	netu---	name
idar-no-in	title/name, Ethernan's	nehtetri	name-?
iddar-no-nn	title/name, Ethernan's	nehht	name, "the father"
iddor-en	name, Idd	nuvvarr	name, Nuffar
ir	numeral 2, second	nehhton	name, "for the father"
<sup>s</sup> irt	death		
		O	
L		<sup>s</sup> onde	adjective large
lietrst	name		
<sup>s</sup> lenab	Bsq. <i>nerhabe</i> child	P	
		+perr	adjective rich
M			
made	verb ? I am ?		

R		RĀTIAN	
rurtes	fem. name, Rurt's	A	
		ais-er	"Gods"
T		ansu-mn-a	devine
tall	name, Tall	ahil	young man? Etr. hil,
tall-u-or	name, Tall		hilar
tall-u-or-ann	name, Tall's	ahua::hukl	
tall-y-on	name, Tall's	akupla	
tied	name ?	alv	
<sup>s</sup> toles	arm, limb	<sup>s</sup> anto	large, Bsq <i>andi</i>
U		apa-nin	mother
<sup>s</sup> usd	Bsq. otso wolf ?	apa-u	father
<sup>+</sup> usc-on-buts	name, "wolf-made youngster"	apa-na	fatherly
		aqir	noun, Etr. <i>apir-as-e</i>
		B	
V		o-balza-na	for the community
vuen-on	uuen? name? monument?	<sup>s</sup> barg	house, cabin
vons	noun ?	<sup>s</sup> bera	rich
viteor	adjective	<sup>s</sup> baites	in the house, at home
vurr	"man"	E	
vorr-enn	name, Vorr's	enikes	name
vurr-act	name, by Vorr	enpetak	?
vrobbacc-en	name, by Frobba's	epetav	?
vvuddaδōs	name, Oddadd	erti	Basque; half
		eris	noun, Etr. <i>eris</i> , ?
		eris-na	adjective ?
		eris-na-ti	adjective ?

t-eris-na	noun	K	
f-eris-na	noun	kali	
esi	Etr. <i>esi</i>	karse	
es-i-mn-es-i		kalun	son ?, name ?
esi-u-mn-i-nusur	?	kaian	name ?
erikian	?	kampelsuri	
estu-ale	this, it ?	<sup>s</sup> karr	noun, stone
etsu-ale	this, it ?	<sup>s</sup> kar-anto	stone-large, height
es stua		karika	building ?
(e)stuva	this, it ?	karapas-na	name
et sua	this, it ?	keliva	name
etain-i	noun ?	kasika-nu	
		kapivap-es	name
H		kalipist-al	name
helanu	?	kastri, kastri-s-ti	noun
		kerrinake	build
I		keriakve	stones
iχ-l	dedicate's	kevi-s-i	name
ipip	here ?	ki	numeral, 3
ilp-si-i, i-ilp-a		-ki	locative suffix
au-ilp-a		<sup>s</sup> klave	sanctuary
istius-na		<sup>s</sup> klav-en-na	TOPONYM,
itt-i, it-u	this ?		sanctuary's
iti-ki-ti		klev-i-e	sanctuary
is-a-e	river	ku	pronoun, thou
is-ar-ki	the river here	kiukui	snake ?
iska, iske, <sup>s</sup> iske	hand, power ?	cicii	
		kolietu	
		knuse	name ?

kulpi-li-na		N	
kunin-as-i	female name	nai	1. pers. pron. sing.
kusen-kus	noun	nakina-tar-i	name, Nakina-from
kus-i-al-e	name or noun	natu-s-i	noun
kus-us	noun	a-nati	the gift
		nati	give
L		nasle	of the dead
lahapru		nlape, nlaup	verb ?
lase-ke, laste	verb	niku	verb ?
lasanu-al-e	name		
las	offer	P	
las-pa, las-pa-s-i	the offering one	-pa	article, one
late-		pal-a, pal-i	cave, grave
laste	plate, altar, tablestone	panaki	
lavisi-e, lavise-al	name	pani-n-i, pani-un	
lemai-s	name	pavis-es	name
lisan-es	name	peke	name
e-luku, e-luku-s	title ?	pel-i, pel-na	
luke, i-luke	companion ?	pelur-i-es-i	name
lup-nu	dead	pelvin-u-al-e	name
la-turu-s-i	the votive ?	perun-i-es	name
		pelipur-i-es-i	name
M		pevasnik-es-I	name ?
malavun		pianu-s	name
maie-ke	verb	pirikanisnu	
<sup>s</sup> meclo	community	pisuauri	
melka	beautiful	pitame	name ?
mu-iu, mu, me	pronoun ?	pitave	name ?
mi	1. pers. pron. sing.	pnake	name

pumis  
prima  
pirima  
puper  
putikinu

## R

<sup>s</sup>rasina      Rätian ethnonym  
remi, remi-k, remi-na  
reitus  
a-rus-na-s, a-ruse-tar      the Rätian  
rusi-e, rus-na-s      man  
rit-a-n, rit-ne, rite      Goddess Rita  
reiti, reite, riti      Goddess Rita  
rit-a-mn-e, reiti-e      Goddess Rita  
riti-e, rit-al-e      Goddess Rita  
ril-ti      lived, age ?

## S

salus  
sakvil-iske      sacred/priest ?  
sakvil,      Ib. *sakar*  
i-sakvil      Etr. *sacni*  
svi      10 ?  
iar-seisvi      year-60 ?  
e-shira-pa      CLASS-star-one  
siara  
siraku  
silnanz      deity Silanus

situ  
skais, sk      made  
<sup>s</sup>sosin, susin-u      fair, just  
spaik  
spura      city, Etr. *spura*  
suprike

## T

taur-iske      mountain-strong  
taukri-li-na      tomb, bury ?  
tepmu  
telra-kinu-a  
ti-kinu-a  
tu-kinu-a  
tina      Etr. *tina*; Etr.  
earthenware pot  
tinu      gift, -u for verbal-  
substantive  
tina-ke      dedicate, (Etr. *zinace*)  
tituli      noun, "name"  
tinia, tine, tines-ma      Goddess Tinia  
tiuti      family, clan  
trahis  
trina-ke      spoke, promised, (Etr.  
*trin*)  
tulie      border, Etr. *Tular*

V		aicensainn
vakhi-kve	bright-plural	aikal
vatanu		--ateins-ai
velkanu	deity Vulkanus	ara-ir
vepelie		asdu-a name
vispeka, vispeka-n		asdu-an-a name
vinu-al-e	of the wine	alisno toponym
vinu-t-al-in-a	of this particular wine	
visia		B
vita-mu		biar
vita-hur		bin
vika		bua-ai
		(bo)an-a-kor-due
U		budu
upiku	verb ?	bue one ?
utiku	verb ?	bue-n-ir
ustitu		bue-n-i-ir-a
u, u-es-i	noun	bue-l-en
uxl		bue-te-i
		bue-r-an-ar
Z		bue-due-to-ar
zekli	wrote. (Etr. zixl)	bue-due-ar-ai
<sup>s</sup> zupar	grave, hidden	bue-mir-n-aR
		bue-kin----
TARTESSIAN		bue-ir-a
A	Comments	bue-ir-o-a
aakaR		bo----i
aibouris..		boue-ir-a
arko		



D		kakosus	name (coin)
durkou-i-ak		kanan, kanan-aR	name
dursen		karkou-i-ak	
duol		ketouibun	toponym, "Ketobriga"
		kirien-a	
E		ko	
eeRe-ir-n-aR		koe	
eeRe-ir-os		kodu	
ekun	Iberian <i>ekuan</i> , NC <i>ekun</i>	koerol-i	
enlin		kon-ai	
eterei	Etr. <i>eteri</i> , "stranger"	kon-aR	verb ?
e(s)a		kon-a-te-i	
eutebi		kon-i	
		kon-i-i	
I		kon-te-i	
ibun	mountain	kon-i-ak	
iro		kon-o	
iro-a		konil	name (coins)
iro-a-s		konil:siskar	name (coins)
iro-an		konisturkis	placename (coin)
iru-al		korani	name (coin)
iirno-st-a			
iu		L	
ilturki	placename (coin)	laanaR	verb ?
iloitvrgens	placename (coin)	lekoe	name ?
		lenku-te-i	
K		loti	
kakkosa	name (coin)		
kakoni	name (coin)		

		S	
M		sa	
mundu	Etr. <i>munth</i> , "netherworld, world"	san-en	
		san-aR	verb ?
mu	pronoun ?	saan-en-aR	verb ?
		san-er	
N		san-o	
natadusates	name ?	san-u	
nine		sako-to-an	
nori-en	Etr. <i>neri</i> , "water"	saro	
narduodu		saro-n-a	
		saro-n-aR	verb ?
O		saro-n-n-aR	verb ?
o-ar-en	Ib. <i>o-ar-en</i>	saro-n-te-i	
oar-la		saro-n-o-te-a	
otagisa	name (coin)	saru-n-o-te-a	
onteo*ta		sati	
ora-o		sati-a	
ore		sate-ir	
otita		siicenii	
		siinkoene	
R		siol-n-aR	
ro		siste:siskra	name (coin)
rola---		sis-te-as	name (coin)
Ratiurs	name ?	sis-k-ar	name (coin)
rato-ak-a		sis-te	name (coin)
resno(n)R/resno(a)R		sisukurhil	name (coin)
		soro-n-aR	verb ?
		sote	

sot-i-u		tio-o	
sot-a-an-a		---tio-re	
sudu	"grave", Etr. <i>suthi</i>	tio-dun-te-i	
sudu-u		tis	
s-tio-o-re		teuinatos	name
T		U	
take	Iberian <i>take</i> , Bsq. <i>-tegi</i> , "here"	ukosaon	personal name
		uurkoe	Ib. <i>urke</i> , "high, fortress"
taati-u		uar	
teir-e-la		uar-a	
teir-o-la		uoin	
teir-due-s-no		uro-a	
tuitis	Celt. <i>tuath</i> , "tribe"	uro-ar	
teu		uro-a-ir	
teo-i		uro-a-ir-s-a	
teo-r-n-aR		utebi	
tio			
tiu			

LITERATURE

- Ambrosini, R.: 1979, *Le iscrizioni sicane, sicule, elime, Le iscrizioni pre-latine in Italia*,  
Accademia Nazionale dei Lincei, Roma, Italia.
- Ammerman, A. and Cavalli-Sforza, L. L.: 1984, *The neolithic transition and the genetics of  
populations in Europe*,  
Princeton University Press, New York.
- Anderson, J.-M.: 1988, *Ancient languages of the hispanic peninsula*  
University Press of America, New York
- Bengtson, J.: 1991, *Dene-Sino-Caucasian languages*,  
In V. Ševoroškin (ed.), *Dene-Sino-Caucasian languages*,  
Universitätsverlag N. Brockmeyer, Bochum, Germany, pp. 67-172.  
Bochum publications in evolutionary semiotics BPX 32.
- Bengtson, J.: 1992, *The Dene-Caucasian macrophylum*,  
In V. Ševoroškin (ed.), *Nostratic, Dene-Caucasian, Austric and Amerind*, Universitätsverlag  
N. Brockmeyer, Bochum, Germany, pp. 334-352.  
Bochum publications in evolutionary semiotics BPX 33.
- Best, J. and Woudhuizen, F.: 1989, *Lost languages from the Mediterranean*,  
Brill, Leiden, Netherlands.  
Publications from the Henri Frankfort Foundation 10.
- Blažek, V.: 1989, *Materials for global etymologies*,  
In V. Ševoroškin (ed.), *Proto-Languages and Proto-Cultures*,  
Universitätsverlag N. Brockmeyer, Bochum, Germany, pp. 37-40.  
Materials from the first International interdisciplinary symposium on language and prehistory at  
Ann Arbor, 8-12 November, 1988.
- Bonfante, L.: 1990, *Etruscan*,  
Trustees of the British Museum, London.
- Bonfante, G. and Bonfante, L.: 1983, *The Etruscan language*  
Manchester University Press, Manchester.
- Cavalli-Sforza, L.L.: 1988, *The Basque population and ancient migrations in Europe*,  
In *Munibe-Antopologia y Arqueologia* Suppl. 6, 129-137.  
San Sebastian, Spain.
- Cavalli-Sforza, L.L., Menozzi, P., and Piazza, A.: 1994, *The history and geography of human genes*.  
Princeton University Press.

- Chamorro, J.G.: 1987, *Survey of the archaeological research on Tartessos*,  
In American Journal of Archaeology 91, 197-232.
- Champion, T., Gamble, C., Shennan, S. and Whittle, A.: 1984, *Prehistoric Europe*,  
Academic Press, London.
- Cristofani, M.: 1979a, *Recent advances in Etruscan epigraphy and language*,  
in D. Ridgway and F. Ridgway (eds), Italy before the Romans,  
Academic Press, London.
- Diakonoff, I.M.: 1985, *On the original home of the speakers of Indo-European*,  
In Journal of Indo-European Studies 13, 92-174.
- Diakonoff, I.M.: 1990, *Language contacts in the Caucasus and the near east*,  
In T.L. Markey and J.A. Grippin (eds), When worlds collide: The Indo-Europeans and the Pre-Indo-Europeans,  
Karoma Publishers Inc., Ann Arbor, pp. 53-66.
- Diakonoff, I.M. and Starostin, S.A.: 1986, *Hurro-Urartian as an Eastern Caucasian language*,  
Muenchener Studien zur sprachwissenschaft.
- Dolgopolsky, A.: 1986, *A probabilistic hypothesis concerning the oldest relationships among the language families in Northern Eurasia*,  
In V. Ševoroshkin and T. Markey (eds), Typology, Relationships and time,  
Karoma, Ann Arbor.
- Forsyth, K.: 1996, *The ogham inscription of Scotland, an edited corpus*  
PhD thesis from the University of Harvard, Cambridge, Massachusetts.
- Gamkrelidze, T.: 1990, *On the problem of an Asiatic original homeland of the proto-indo-europeans*,  
In T.L. Markey and J.A. Grippin (eds), When worlds collide: The Indo-Europeans and the Pre-Indo-Europeans,  
Karoma Publishers Inc., Ann Arbor, pp. 5-14.
- Greenberg, J.: 1987, *Language in America*,  
Stanford University Press, San Francisco.
- Hewitt, B.: 1981, *Caucasian languages*,  
In B. Comrie (ed.), Languages of the Soviet Union,  
Cambridge, Cambridge.
- Hubschmid, J.: 1960, *Mediterranean substrate*,  
In Romanica Helvetica 70, 4-97.

- Hubschmied, J.: 1982, *Vorindogermanische und indogermanische substratwörter in den romanischen sprachen*,  
In S.Ureland (ed.), Die leistung der substratforschung,  
Max Niemeyer, Tübingen.
- Illič-Svityč, V.M.: 1989, *The relationship of the Nostratic family of languages: A probabilistic evaluation of the similarities question*,  
In V.Ševoroškin (ed.), Explorations in language macrofamilies,  
Universitätsverlag N. Brockmeyer, Bochum, Germany, pp. 111-113.  
Materials from the first International interdisciplinary symposium on language and prehistory at  
Ann Arbor, 8-12 November, 1988.
- Jackson, K.~H.: 1955, *The Pictish language*,  
In F.T. Wainwright (ed.), The problem of the Picts,  
Nelson, London.
- Kaiser, M. and Ševoroškin, V.: 1987, *On recent comparisons between language families; The case of Indo-European and Afro-Asiatic*,  
In General linguistics 27, 34-46.
- Kammenhuber, A.: 1975, The linguistic situation of the 2nd millennium BC in ancient Anatolia,  
In Journal of the Royal Asiatic Society 87, 116-120.
- Kretschmer, P.: 1943, *Die Vorgriechischen Sprach und Volksschichten*,  
In Glotta 30, 84-217.
- Locker, E.: 1962, *Die ältesten sprachschichten westeuropas*,  
In Österreichische akademie der wissenschaften, philosophische-historische klasse, sitzungsberichte 240, 5-59.
- MacAllister, R. A.S.: 1940, *The inscriptions and language of the Picts*,  
In J. Ryan (ed.), Essays and studies presented to Professor Eoin MacNeill,  
Clith, Dublin.
- McAlpin, D.W.: 1981, *Proto-Elamo-Dravidian: The evidence and its implications*,  
In Transactions of the American Philosophical Society 71:3, 1-155.
- Nikolaev, S. and Starostin, S.: 1991, *North Caucasian roots*,  
In V. Ševoroškin (ed.), Dene-Sino-Caucasian Languages,  
Universitätsverlag Dr. Norbert Brockmeyer, Bochum.
- Oroz-Arizcuren, F.J.: 1981, *La relation entre el Vasco y el Iberico desde el punto de vista de la teoria del sustrato*,  
In Iker; Encuentros Internacionales de Vascologos, Bilbao 1, 241-255.

- Pallotino, M.: 1975, *The Etruscan*,  
Allen Lane, Penguin Books, London.
- Pallotino, M.: 1987, *Proposte, miraggi, perplessità nella ricostruzione della Storia Etrusca*  
In Studi Etrusci 53, 487-522
- Pattison, W.: 1981, *Iberian and Basque--A morpho-syntactic comparison*,  
In *Archivo de Prehistoria Levantina* 16, 487-522.
- Pedersen, H.: 1933, *Zur frage nach der urverwandschaft des Indo-Europäischen mit dem Finno-Ugrischen*,  
In Memoires de la societe finno-ougrienne 67, 308-325.
- Pittau, M.: 1981, *La lingua dei sardi nuragici e degli etrusci*,  
Editrice libreria dessi, Sassari, Italia.
- Renfrew, C.: 1989a, *Archaeology and language-The puzzle of Indo-European origins*,  
Penguin Books, 27 Wrights Lane, London W8 5TZ, Great Britain.
- Renfrew, C.: 1994a, *Before Babel, Speculations on the origins of linguistic diversity*,  
In Cambridge Archaeological Journal 1, 3-23.
- Rhys, J.: 1892, *The inscriptions and language of the northern Picts*,  
In Proceedings of the Scottish Archaeological Society 26, 263-351.
- Rhys, J.: 1898, *A revised account of the inscriptions and language of the northern Picts*,  
In *Proceedings of the Scottish Archaeological Society* 32, 324-398.
- Rix, H.: 1968, *Eine morphosyntaktische uebereinstimmung zwischen Etruskisch und Lemnisch; die datierungsformel*.  
In Studien zur Sprachwissenschaft und kulturkunde. Gedenkschrift W. Brandenstein,  
Germany, pp. 213-222.
- Rix, H.: 1985, *Schrift und Sprache*,  
In M. Cristofani (ed.), Die Etrusker,  
Beiser Verlag, Stuttgart, Germany, pp. 210-238.
- Rix, H.: 1991, *Etruskische texte, Band I+II*,  
ScriptOralia, Gunter Narr Verlag, Tuebingen, Germany.
- Roman del Cerro, J.L.: 1993, *El origien iberico de la lengua vasca* ,  
Editorial Aguacalra, Alicante.
- Ruhlen, M.: 1994, *On the origin of languages*,  
Stanford University, Press, Stanford.

- Scharf, H.: 1978, *Goethes Morphologiedefinition und das Problem der Cromagniden zu den Urgermanen*  
In Gegenbaurs Morph. Jahrbuch 124, 139-190.
- Scharf, H.: 1982, *Linguisitische Vergleiche zur Ergänzung morphologisch-antropologischer Untersuchungen*,  
In Verh. Anat. Ges. 76, 557-566.
- Schmidt, K.H.: 1985, *A contribution to the identification of Lusitanian*,  
In J. Hoz (ed.), *Actas del III coloquio sobre lenguas y culturas paleohispanicas*,  
Universidad de Salamanca, Spain.
- Schmoll, U.: 1961, *Die Südlusitanischen Inschriften*,  
Harassowitz, Wiesbaden.
- Schuhmacher, S.: 1992, *Die Rätischen Inschriften, Geschichte und heutiger stand der Forschung*,  
In Innsbrucker beiträge zur kulturwissenschaft, Sonderheft 79  
ISBN 3-85124-155-X, Innsbruck.
- Starostin, S.: 1989, *Nostratic and Sino-Caucasian*,  
In V. Ševoroškin (ed.), Explorations in language macrofamilies,  
Universitätsverlag N. Brockmeyer, Bochum, Germany, 42-66. Materials from the first  
International interdisciplinary symposium on language and prehistory at Ann Arbor, 8-12  
November, 1988.
- Stoltenberg, H.L.: 1943, *Die Bedeutung der Etruskischen Zahlnamen*,  
In Glotta, 30, 234-243.
- Sverdrup, H.: 1995, *Classifying and translating inscriptions in the Pictish language*,  
In Reports in ecology and environmental engineering 1995:4.
- Sverdrup, H.: 1997, *A typological and morpho-syntactical analysis of the extinct Rätian language*,  
In Reports in ecology and environmental engineering 1997:3.
- Sverdrup, H.: 1999, *Ecological modelling of language origins, relatedness and parths of dispersal*;  
*Part I; The LANGUAGE model*.  
To be published in the series Reports in ecology and environmental engineering during 1999.
- Sverdrup, H. and Guardans, R.: 1998a, *The Swedish-Spanish Iberian language study 1; Analysis of  
some utilitarian texts in the ancient Iberian language*.  
Manuscript prepared for submission during 1999.
- Sverdrup, H. and Guardans, R.: 1999a, *The Swedish-Spanish Iberian language study 2; Analysis and  
attempted interpretation of the lead from Alcoy*.  
Manuscript prepared for submission during 1999.



- Sverdrup, H. and Guardans, R.: 1999b, *The Swedish-Spanish Iberian language study 3 : A morpho-syntactical study of Iberian texts from Solaig, Emporion, Castellon and Ullastret*.  
Manuscript prepared for submission during 1999.
- Sverdrup, H. and Guardans, R.: 1999c, *The Swedish-Spanish Iberian language study 4: Genetic relationships and typology of the ancient Iberian language*.  
Manuscript prepared for submission during 1999.
- Sverdrup, H. and Guardans, R.: 1999d, *Investigating Tartessian inscriptions for a consistent transcription and an exploratory study of the language*,  
Manuscript prepared for submission during 1999 .
- Tovar, A.: 1951, *Lexico de las inscripciones Ibericas*,  
In R. Mendez-Pidal (ed.), Estudios dedicados a Mendez Pidal,  
Madrid.
- Tovar, A.: 1961, *The ancient languages of Spain and Portugal*,  
S. F. Vanni Publishers and Booksellers, New York.
- Tovar, A.: 1977, *Krahes alteuropäische hydronomie und die westindogermanische sprachen*,  
Carl Winter Universitätsverlag, Heidelberg.
- Untermann, J.: 1975-1990, *Monumenta Linguarum Hispanicum* 1975; I--Die Munzlegenden, 1980;  
II--Die Iberischen inschriften aus Frankreich, 1990a; III Die Iberischen inschriften aus  
Spanien, Literaturverzeichnis, einleitung, indices, 1990a; IIb, Die inschriften.  
Dr. Ludwig Reichert Verlag, Wiesbaden
- Vennemann, T.: 1994, *Linguistic reconstruction in the context of European prehistory*,  
In Trans. Phil. Soc. 92, 215-284.
- Villar, F.: 1990, *Indo-européens et pre-indo-européens dans la péninsule Iberique*,  
In T.L. Markey and J.A. Grippin (eds), When worlds collide: The Indo-Europeans and the Pre-Indo-Europeans,  
Karoma Publishers Inc., Ann Arbor, pp. 363-394.
- Woudhuizen, F.: 1992a, *The language of the sea peoples*,  
In Publications from the Henri Frankfort Foundation 12.  
Najades Press, Amsterdam, Netherlands.
- Woudhuizen, F.: 1992b, *Linguistica Tyrrhenica*,  
J. C. Geben, Amsterdam, Netherlands.



# FAMILY EVOLUTION, LANGUAGE HISTORY AND GENETIC CLASSIFICATION<sup>1</sup>

Ilia Peiros

Lexicostatistics<sup>2</sup> remains a pariah of comparative linguistics. Sometimes linguists agree to use it as very preliminary tool of investigation, but more often they totally reject it on the basis of general observations. No present day discussion of lexicostatistics has been conducted so far.

When discussing lexicostatistics, we need to address the following issues:

- What are the methods used in lexicostatistics?
- Why is it used?
- How can it be used?
- How does it fit into the broader framework of comparative linguistics?
- How does a lexicostatistical classification of a family correspond to other classifications of the same family proposed used in comparative linguistics?

The theory of lexicostatistics is the subject of a monograph entitled 'Lexicostatistics revisited' which Starostin and I are working on. Currently the structure of the monograph is the following:

Chapter I. The theory of lexicostatistics.

Chapter II. Glottochronology.

Chapter III. Lexicostatistics as a heuristic.

Chapter IV. Case Studies.

Here I present some parts from Chapter I, as they are written by me in Melbourne in 1997 on the basis of discussions with Starostin which started about twenty years ago and resume every time we see each other. Unfortunately, the distance between Melbourne and Moscow (I hope only a geographical one!) did not allow me to finalise this text with Starostin and to present it as a joint publication.

---

<sup>1</sup> I am grateful to Marc Durie for his help and inspiration. The final stage of this project has been supported by an Arts Faculty Completion Grant of the University of Melbourne.

<sup>2</sup> We distinguish here lexicostatistics (a method of genetic classification of languages) and glottochronology (a method of obtaining absolute datings in comparative linguistics).

Therefore I am publishing it under my name, hoping, however, that the article correctly reflects the essence of the future monograph.

§1. Any human community can be represented as an informational network which facilitates direct or indirect communication between its members. Such a reliable permanent informational network connecting all its members is probably the most important condition of the very existence of a community: if communication is not maintained, a community has little chance to survive.

The media of communication is always human languages which form a linguistic repertoire of a community (a set of languages used in it). This repertoire is not necessarily limited to one language, and multilingual communities are not less typical than monolingual, but the necessity of reliable informational exchange is normally based on the fact of shared language knowledge<sup>3</sup>.

This leads us to the well-known observation that the main function of any natural language is to maintain informational exchange between people who belong to the same community.

Each community has certain views about languages used by its members and in other communities. These linguistic views are not necessarily identical to the views accepted by professional linguists. Therefore we will distinguish between the two notions, 'language' and 'sociolanguage'. Although two speakers of the same language can use it with sometimes quite noticeable differences, when discussing common topics they will always understand each other. This makes the criterion of mutual intelligibility to be essential for the notion 'language'<sup>4</sup>. Two speech varieties belong to one sociolanguage if the speakers believe that they speak the same (socio)language, regardless of their actual ability to understand each other. Languages and sociolanguages form different combinations:

#### One language - one sociolanguage

Speakers of Hungarian know that they use the same (socio)language. Minor differences do not prevent mutual intelligibility, which means that they belong to the same language.

#### One language - two or more sociolanguages

<sup>3</sup> This is not true for more complex entities of human organisation, for example empires, like British or Roman.

<sup>4</sup> Dialectal chains, if they exist, do not contradict this claim (Peiros 1989)

This situation is represented, for example, by Serbian and Croatian: mutual intelligibility (one language), but people know that they speak different (socio)languages.

Several languages - one sociolanguage.

The Chinese 'dialects' present an excellent example of this type of relation. It is well known that the differences between some of them are not less than between major Slavonic languages (where independent language status is generally accepted) and mutual intelligibility is normally not possible. At the same time speakers of Chinese 'dialects' know that they speak the same (socio)language and are prepared to defend this claim.

In comparative linguistics the main focus is on languages, not sociolanguages. Therefore a language family is seen as formed only by languages, where differences can be evaluated by formal methods without an appeal to the views of their speakers.

§2. Any traditional community lives in its habitual world with a well-known environment, usual activities, customary social relations and other predictable features. In such a life, similar events usually occur more or less frequently and it is always known how a member of the community is supposed to act, what should be said and what kind of response is expected from other people involved. This knowledge and the ability of predictions are essential for the well being of any community.

Using linguistic tools, members of a community are able to convey any type of information regarding their everyday activities, typical situations, environment and so on. It is still to be investigated how complete and precise is this information, but from the theoretical point of view it is clear, that all basic informational demands of a stable traditional community are met by language(s) used in it. It does not mean, however, that any type of information can be easily communicated in any language, as often speakers have trouble when trying to express alien ideas in their own language. In any typical situation, however, each community has sufficient linguistic means to communicate appropriate information. This information is always community-specific and it is practically impossible, for example, to use Russian when talking about trees of the Indian jungles or to discuss ancestor rituals of Hmong in English.

The predictable way of life can be maintained while the community lives in the world with no significant changes: it occupies the same territory, it is not forced to change its activities, its neighbours and cultures also remain the same. A migration to a new environment or contacts with new cultures would undermine stability and the community would face the problem of adaptation to previously unknown situations. Such an adaptation always affects the existing linguistic repertoire. The community can either adopt a new language which is better prepared for the new life, or adjust the old one to new demands.

§3. Changes in the life-style are usually related to adaptation of new cultural ideas: new objects of material culture, new skills and views and if the community migrated, knowledge of the new environment. Over a certain period of time these changes would be reflected in the community's language(s).

It is generally believed that the culture of a community is represented in its language, mainly in its cultural lexicon, which is formed by words related to various cultural ideas. New ideas are usually represented either by old words with modified meanings, or by borrowings, as it is quite common for people to borrow ideas together with the appropriate labels (words) (see, for example, Simpson 1985). Therefore, the cultural lexicon of a language

- is historically not very stable and can be significantly changed over a short period of time;
- can include many borrowings reflecting the process of cultural adaptation.

A migration to a new territory with unknown vegetation and animals can also cause significant changes in the so-called 'environmental' lexicon of the language, which is formed by various words with meanings related to the natural world. Under certain circumstances we can expect to find here quite significant changes reflecting the differences of the two territories. In some cases the reorganization of 'environmental' lexicon is achieved mainly by changing meanings of original words and forming new complex expressions (see, for example, Biggs 1991). In other cases the main emphasis is put on loans.

There is, however, a particular part of a lexicon, which is less affected by changes in the community's life. We can identify more or less a universal set of ideas known in all or nearly all human communities regardless of their level of cultural development, territories occupied and other properties. These ideas are quite

basic and simple: 'moon', 'sun', 'man', 'water', and many others. No doubt that languages differ considerably in the ways they present these ideas, but for any language, there are always means to represent them, unlike environmental or cultural ideas. Therefore it seems useful to talk about 'core information' and **core meanings** representing it. These core meanings can be represented in languages in many different ways, but it seems to be possible to identify such meanings across languages. As there are no obvious or universal reasons why core meanings should be changed in time, they are generally more stable than, say, environmental or cultural meanings and normally are better preserved in languages.

§4. Any human community *N* always has at least one language *T* known to its members. This language can be either inherited from the previous generations of speakers or be learned through contacts with other people. There are no recorded cases of glottogenesis<sup>5</sup> - the creation of a new language in total isolation from other languages and apart from the earliest periods of human history (not studied by comparative linguistics), glottogenesis is not possible.

The relation between community *N* and its language *T* always reflects the current status of *N*: whether it is stable or is in the process of formation or disintegration.

Over the whole period of stability, any two consecutive generations of *N* share the same language *T* learned from their parents and are able to communicate using it. No other languages are acquired by *N* in this period and its linguistic repertoire remains unchanged. The norms of language usage accepted by the majority of the community's members govern communication and sanction all changes to the system of *T*.

The period of formation of a new community is actually the period of establishing a new informational network to connect all its potential members. Among the important considerations here are:

- the necessity to have a reliable network functioning with minimum distortions;
- the necessity to have this network operational in the shortest possible time.  
(The longer the period, the lesser the chances are for bringing together potential members),

---

<sup>5</sup> Formations of pidgins and creoles always take place in the situation of linguistic contact.

- the necessity to have a user-friendly network, especially suitable for potential members in more prominent social positions.

All these considerations indicate that the only acceptable strategy in creating such a network is an adoption (at least as the basis for communication) of a language, already used for such purposes.

A new community can be formed either through disintegration of an earlier community or through crystallization from previously distinctive human groups.

It is important, however, to remember that in both options, languages are always either inherited or borrowed.

The process of a community's disintegration is related primarily to the collapse of a single system of norms accepted by all its members of the original community. Due to various extra-linguistic reasons (migration, political turmoils, etc.), norms adopted in community *N* lose their authority over potential members leading towards disintegration of *N* into several daughter groups. At very early stages of disintegration, norms governing the behavior of members are more or less identical across these groups, reflecting their common origin. Later, each group begins to develop its own norms no longer following a common pattern. This can be caused by difficulties in maintaining communication (the groups are not in contact any more), or is the result of purposeful attempts to create and maintain a new identity ('We should not dress like them', 'We do not say this word', etc.). Over a certain period of time, the accumulated effect of changes caused by the distinctive sets of norms would wipe away most of the features of common origin.

When a new community is crystallized from previously different groups it either adopts a language used by one of these groups, or borrows one from neighbours. Both options are explained by the following model. In any stable human group one can always identify a dominant subgroup which is also the center of the informational exchange. The language of this dominant group will inevitably be used for communication. If another language is chosen, it would mean the loss of influence from a previously dominant group (it would be in a disadvantaged position in communication) and the rise of another dominant group (associated with the chosen language). The dominant group can be a part of a new community, or it can exist separately from it, as it happens in the formations of creoles. In both cases, however, a language, which is tightly associated with power, is adopted as the basis of communication.



Therefore a crystallizing community acquires its language:

- (i) either by inheriting the language of its ancestors (of the whole new community) or some of them (of a certain part of it); or
- (ii) through borrowing from another community.

Under no conditions could glottogenesis take place.

From the above discussion it follows that we need to distinguish between the history of language *L* (changes in its system: phonology, morphology, lexicon, and so on) and the linguistic history of a community: its maintenance or language shift<sup>6</sup>. There are three main options in the linguistic history of community *N* with language *T*:

- (i) *N* has inherited *T* from previous stages of its development;
- (ii) *N* has borrowed *T* from another community *T*, either adding it to its repertoire, or using *T* instead of its original language (a shift to *T*);
- (iii) *N* has stopped using *T* for any type of communication.

In the cases of (i) and of (ii) *T* would possibly undergo many significant changes and thus it would be rather different from the original one, but under no circumstances would a new language (not based on a language or languages, which already exist) be invented by *N*.

An uninterrupted development of *T* is related to options (i) and (ii), reflecting two possibilities of language acquisition by its speakers: through inheritance or borrowing.

§5. Every spoken language is the subject of a permanent process of changes, based on the high redundancy of human languages.

Change  $\varphi < \alpha, \beta, Q, t >$  is a process which occurs in the system of a language and can affect any element of it:

$\alpha$  is the initial stage of change  $\varphi$ ;

$\beta$  is the outcome of change  $\varphi$ ;

$Q$  is a set of conditions under which change  $\varphi$  took place;

---

<sup>6</sup> If in the process of community development language *A* is supplanted by a borrowed language *B*, we will talk about a shift from *A* to *B*.

t is a period of time when change  $\phi$  took place.

Depending on the relation between  $\alpha$  and  $\beta$ , we can distinguish:

	Period A		Period B
drifts	$\alpha$	$\Rightarrow$	$\beta$
losses	$\alpha$	$\Rightarrow$	$\alpha$
additions	$\alpha$	$\Rightarrow$	$\beta$

In a drift,  $\beta$  is an interrupted development of  $\alpha$  and is its reflex.

A change can be caused externally or internally. An external causation of a change can be due to the influence of another language, or it can reflect a conversion of several languages within a linguistic area. However, in real practice it is sometimes difficult or even impossible to determine the true nature of the causation.

Changes in a language can either be triggered or free. If a change is caused by another change, which happened earlier, we are dealing with a triggered change. Otherwise a change is a 'free' one. Extralinguistic features, which are often behind the linguistic changes (especially in the lexicon), are not seen here as triggers. Therefore, if a new word is created to represent a new idea, this new idea is not seen as a triggered change. But a split of vowels which took place as a result of development of register is seen as a triggered change.

Altogether we can identify 12 different types of changes:

	Internal		External	
	Triggered	Free	Triggered	Free
drifts	1	2	7	8
additions	3	4	9	10
losses	5	6	11	12

### 1. Triggered internal drift

For example: the change of meaning of the English word 'hound' in the process of the adoption of the word 'dog';

### 2. Free internal drift

For example: the retention of Old English words in their modified modern forms;

3. Triggered internal additions

For example: the development of the fixed word order caused by losses of various morphological distinctions;

4. Free internal additions

For example: the creation of new compounds to represent new ideas;

5. Triggered internal loss

For example: the loss of a distinction between two noun cases, caused by the loss of final vowels;

6. Free internal loss

For example: the loss of words - labels of artefacts not used anymore;

7. Triggered external drift

For example: the development of articles under the influence of another language;

8. Free external drift

For example: the usage 'Ja ne dumaju' (literally 'I don't think') instead of 'Ja v etom ne uveren' to map the English expression 'I don't think so' by Russian migrants in the English speaking world;

9. Triggered external addition

For example: the development of classifiers in many Southeast Asian languages (the actual forms can develop due to internal drift)

10. Free external addition

For example: a lexical borrowing as a label for a new concept;

11. Triggered external loss

For example: a loss of first syllables in many Southeast Asian languages (e.g., Vietnamese or Chamic) as a result of regional convergence;

12. Free external loss

For example: a displacement, when an original word is lost and its functions are now performed by a borrowing.

These types of changes can be found in the history of any language whose development is, in fact, the process of accumulations of changes' outcomes.

§6. Studying the history of language *L* we need to identify at least:

(i) individual changes, which took place in the process of *L* formation and afterwards. This study always includes a description of the four components ( $\langle \alpha, \beta, Q, \tau \rangle$ ) of these changes;

(ii) pairs of triggers and triggered changes;

(iii) relative chronology of the changes.

In such a study, we are supposed to investigate both external and internal caused changes and if possible, to specify the sources of their causations. In many cases, however, this cannot be done. If, for example, the language which was a source of intensive borrowings for *L* is not known, we often cannot identify the borrowed words in *L*. This, however, would not prevent us from describing the history of *L*.

Every single language has its own linguistic history which is always different from the history of any other language, because the changes and their chronology are always language specific.

The accumulation of various changes in a language's system is the process of language development. Changes affect all elements of a language system with no exceptions: nothing in a language is immune to change.

Over a certain period of time the language used by descendants of group *N* could become quite different from language *T* once spoken by *N*, even without a language shift. Accumulated substitutions can wipe away original features of *T*, leaving us with a question of how we can demonstrate that language *L* is a development of *T* and not, say, *R*.

This leads us to the notion of **language continuity**. We will talk about language continuity from period *P* to period *P'*, if for the whole time *t* elapsed between these periods, for every consecutive pair of generations of speakers, the core information was conveyed mainly with the help of linguistic expressive means of the same origin. In other words, language *L* is a continuation of *T* if its expressive means for core meanings have been inherited from *T* and are mainly the result of various drifts, rather than additions.

What does it mean, however, 'mainly with the help of linguistic means of the same origin'? The formal answer would be that if 51% of such meanings in language A' came from language A and 49% came from B, we talk about language continuity from A to A'. If later, the balance would change and some more linguistic meanings from B would substitute meanings from A, we would have to assume that a language shift took place and A' is now a continuation of B, rather than A'. It does not mean, however, that we accept the idea that a genetic affiliation of a language can be changed in time. We have only registered a language shift: before the change of that balance, people spoke a language which was a continuation of A which was full of borrowings from B, while after the shift they began to use a continuation of B with borrowings from A.

§7. Language *T* is often subject to the process of disintegration accompanied by the formation of two or more new languages, each being a language continuity of *T*. The following model explains this mechanism:

In a certain period, there was a community *T* associated with language *T*. The usage of *T* was governed by an extremely complex set of norms which were more or less obligatory for all members of the community. The development of *T* was also governed by these norms, and only the changes approved by these norms were incorporated into the language system.

Due to various extra-linguistic reasons, community *T* began to disintegrate into several groups which supplanted *T*. At the very early stages of disintegration, the norms of language usage were more or less identical for all these groups, reflecting their common origin. Later, the groups began to adopt norms, which were not necessarily shared by all of them. These different norms sanctioned different changes to the original identical language system. Their accumulated results caused a split of a previously common language into its daughter-languages L, L', L'', each having a language continuity from *T*. Over time, these new languages began in their turn to disintegrate following the same model as their ancestors. This led to a formation of a language family which includes all language continuities from *T*.

---

<sup>7</sup> It is worthwhile to mention that no recorded cases of such situations are known to us.

It follows from the previous discussion that if  $L$  is a result of development of  $T$ , under no circumstances would it become a development of  $T'$  which is not a continuation of  $T$ . Obviously, hypotheses about genetic affiliation of  $L$  can be changed, but not the affiliation itself. A community can also change its language, and it is highly possible that its descendants would use a language of another genetic affiliation. In such cases we are dealing, however, with a language shift, rather than with a shift of genetic affiliation: a position of a language in a language family is permanent and does not change in time.

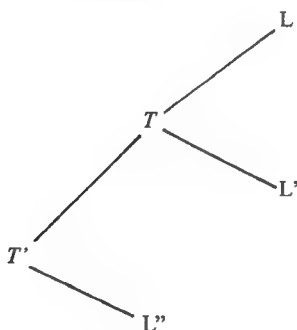
Let us introduce some more technical terms:

If  $L$  is a language continuation of  $T$  we will call  $L$  a 'descendant' of language  $T$ , while  $T$  is the ancestor of  $L$ .

Language  $L$  is a **daughter-language** of  $T$  if:

- (i) there is language continuity from  $T$  to  $L$ ; and
- (ii) there is no such  $T'$  which is an ancestor of  $L$  and a daughter-language of  $T$ . In Figure I all five languages are in the family of  $T$ , but only four of them are related directly:  $T/L$ ,  $T/L'$ ,  $T'/L''$  and  $T'/T$ .

Figure I



All languages with a common ancestor form a **language family**. Languages which belong to the same language family are genetically related. In other words, if two languages  $L$  and  $L'$  are genetically related they always have a common ancestor  $T$  and they are uninterrupted continuations of  $T$  in time. Strictly speaking, therefore, if we want to demonstrate that the languages are related, we should be able to present their common ancestor.

Two languages are specifically related if they are daughter-languages of the same ancestor language.  $L$  and  $L'$  in Figure I are specifically related, as are  $T$  and  $L''$ . Languages  $L$  and  $L''$  are not specifically related as they have different direct ancestors. Specifically related languages, let us call them **sister-languages**, form separated groups within the family, sometimes called 'branches' of the family.

§8. When comparing languages  $T$  and  $L$  which are separated by a significant period of time, one needs to address two logically independent questions:

- is there language continuity between  $T$  and  $L$ ?
- is feature  $\beta$  found in  $L$  a reflex of feature  $\alpha$  found in  $T$ ?

Answers to these questions are not necessarily interrelated and often, despite language continuity between  $T$  and  $L$ ,  $\beta$  is not a development of  $\alpha$  as it is for example a borrowing from an unrelated language. Usually we do not have much trouble interpreting such situations, and we simply say that  $T$  and  $L$  are genetically related but that  $J$  is a borrowing. However, in the literature, one can find discussions of more complex situations of mixed languages, namely languages which have had more or less equal amounts of elements of different origin, such as say, in  $L$ , where the source of its lexicon is language  $A$  and the source of its morphology is language  $B$ . The question typically asked for such situations is: how can we identify the genetic affiliation of this language - as a development of  $A$  or of  $B$ ? According to the definition given above, the language continuity is determined by linguistic means associated with core meanings, and so the language which is the main source of such meanings in  $L$  should be recognized as its ancestor. Unfortunately we have no access to detailed descriptions of mixed languages, so we cannot support this theoretical observation by an investigation of a real case.

The internal development of language  $T$  is caused by various changes which affected all parts of its system and, after a certain period of time, accumulated results of such changes make the new system so different from the original one that

we have to talk about a new language *L* which is different from *T*. It is important to mention that a particular change never affects language development on its own, and only the accumulation of many changes causes a language disintegration.

The process of internal development of *L* can be represented as consisting of the following major stages:

stage I is *L*'s crystallization: due to the accumulation of specific changes *L*'s system becomes significantly different from the systems of its sister-languages;

stage II is *L*'s internal evolution when *L*'s system remains relatively homogeneous. As the changes of *L*'s system continue to be accumulated, the differences between *L* and its sister-languages are constantly increasing;

stage III is *L*'s disappearance caused either by its disintegration or by its 'death' (*L* or its daughter-languages are not used any more).

These three major stages of language development are not separated from each other by any sharp dividing line. To the contrary, in most cases, the transition from one stage to another is rather gradual and takes a reasonably long period of time. For example, if we talk about language disappearance due to its disintegration (the transition from stage II to stage III) we can expect that at first the differences between *A*, *A'* and *A''* (which would later develop into *L*'s daughter-languages) were minimal (if any) and most changes were common to all of them, while later the changes would become language-specific. It is highly probable that speakers of *A*, for example, used forms of *A'* and did not identify them as foreign additions to their own language. Such borrowings often cannot be detected by comparative methods. That is why we have to assume that each stage is separated from another by a certain 'blind spot' whose features and duration are not quite known.

§9. The history of a language family is formed by the histories of individual languages. First its common proto-language goes through the three major stages of evolution (crystallization, integrated development and disappearance), then its daughter-languages go through them, then in turn the daughter-languages of these descendants, and so on. As a language always has only one ancestor and cannot



change it in time, this aspect of family evolution (its branching) has to be represented by a genetic tree of a fixed structure.

We distinguish between the notions of 'genetic tree' and 'evolutional tree'. A genetic tree is only a presentation of a family structure, while an evolutional tree is of a more complex nature to be discussed below.

As a genetic tree represents only linguistic continuity, for each language *L* we need to know only:

- (i) *L*'s direct ancestor (if any);
- (ii) *L*'s sister languages (if any);
- (iii) *L*'s daughter-languages (if any).

The following formal features of genetic trees reflect our understanding of a family's evolution:

1. A genetic tree represents the internal structure of a family and thus includes only genetically related languages.
2. A genetic tree is formed only by nodes and directed arcs connecting these nodes.
3. There are two types of nodes:
  - (i) those representing recorded languages;
  - (ii) those representing entities postulated for the needs of classification.

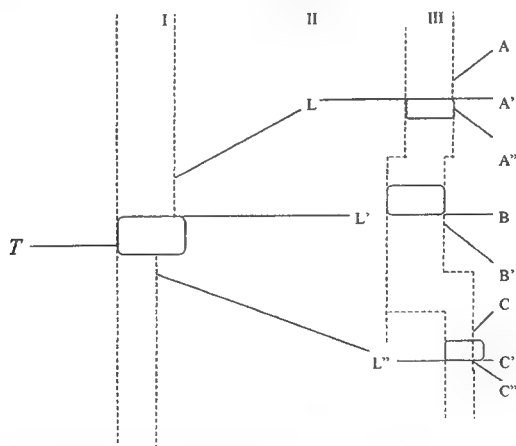
Often, but not always, these entities represent reconstructed languages postulated in the process of the family's investigation.
4. A directed arc represents the linguistic community connecting ancestor language *T* with its daughter-language *L*.
5. In each tree there is only one root node with no entering arc corresponding to the common proto-language of the family. This reflects the assumption that all related languages have developed from one common source - the proto-language of the family.
6. Apart from the proto language, every other language of the family always has only one direct genetic ancestor. In a genetic tree, every node, other than the root one, has only one entering arc. No node can have more than one entering arc.
7. A node without any arcs going out represents a language without known descendants; it could either be a language with no speakers (= a dead language) or a language without a significant dialectal diversity.

8. There is no limitation on the number of arcs going out of a node: none, one, two, or many. No well grounded reasons are known to support the idea that a genetic tree should always have a binary structure<sup>8</sup>.
9. If two nodes are connected by an arc, this connection remains unchanged and under no circumstances do we accept a situation when, at one chronological stage, a node is connected to one ancestor, while in a later period, it is connected to another one. This reflects our fundamental belief that a language never changes its genetic affiliation.

It follows from the above that in order to create a genetic tree, we need to define how to identify nodes and how to connect them with arcs. As differences between nodes are caused by differences of accumulated changes rather than individual changes, there is no logical necessity to investigate these particular changes individually.<sup>9</sup>

§10. Each stage of language development has its own duration which is not necessarily identical across the family, and at any given period of time related languages can represent various stages of development (see Figure II).

Figure II



<sup>8</sup> Binary branchings in classifications often reflect not the internal structures of families, but rather the properties of classification procedures.

<sup>9</sup> This claim contradicts the generally adopted technique of classification with the help of innovations.

In this Figure Stage I covers the disintegration of ancestor language *T* and the formation of its daughter-languages (*L*, *L'*, *L''*). In this period, the daughter-languages start to develop independently, and all common features adopted by them during and after this period are caused by borrowings, convergences, and other non-genetic reasons. Stage II covers the periods of integrated independent development of languages *L*, *L'* and *L''*. Stage III covers the period of disintegration of the daughter-languages: *L* develops into *A*, *A'* and *A''*, *L'* develops into *B* and *B'* and so on.

§11. Various changes in a language do not occur spontaneously, and the following explanation can be suggested.

Knowledge of a language is, first of all, an ability to express a given meaning in a text (verbal or written) and to extract all possible meanings from a given text. From this point of view, language is a device which matches meanings and texts (Mel'chuk 1988, p. 47). Speakers, however, also know the norms of the language usage adopted in their community. These norms allow them to make choices within options provided by the language: how to choose one of the synonyms, how to pronounce a particular sound, which morphological or syntactic structure is preferable in this situation, and so on. These norms are not necessary identical for all its members (cf. sociolinguistic variations within a community), but if a member speaks 'properly', he or she always knows the norms and follows them.

Different communities which use the same language do not normally share the sets of such norms and their members are often aware of that ('We do not use this word', 'This is the American pronunciation', and so on).

All changes in a language are always triggered by changes in the norms of usage. At first, a new norm only suggests the preferable choice among several options provided by the language, but after a certain period of time, this choice becomes the only acceptable possibility which leads towards changes of the corresponding linguistic structures.

Norms of language usage are in a constant process of change which in turn causes changes in the language structure, which accumulate such changes and results in significant differences between the various stages of the language development.

The acceptance of norms and their changes is crucial for language development. If various groups within the community began to adopt different

norms this could lead towards the development of differences in the language they use: different norms trigger different linguistic changes and eventually would end up with the language's disintegration and formation of several new languages, each being a daughter-language of the same original one.

Therefore we can postulate the following connection:

change of norms of usage => changes of language's features => family evolution.

Of these, only family evolution is of a universal nature, while change of norms are community specific and changes of a language's features are language specific.

Therefore we have to assume that development of a language family is not just a simple sum of histories of individual languages. It can be seen as formed by three different processes:

1. Family evolution, which is a universal process of branching: an original language disintegrates into new ones, each being its daughter-languages.
2. Histories of individual languages formed by numerous changes affected systems of these languages. These changes are usually language-specific, but sometimes similar changes can be found in different languages. It is important to mention that a particular change in a language's history does not affect the family evolution. The statement, 'if a change  $*a > \beta$  did not occur we cannot talk about the split of languages A and B', is not correct.
3. History of norms of language usage within a particular community are always community-specific.

Here we will discuss only problems related to a family's evolution.

§12. When discussing a genetic classification, linguists usually distinguish two problems:

- (i) identifying the genetic affiliation of languages: Germanic, Sino-Tibetan, Nostratic, etc.
- (ii) creating an evolutionary tree for a family.

When working out an evolutionary tree for a family, we must first identify which languages are genetically related and form this family. To do that one can use the following formal procedure.

Two languages A and B are genetically related if the three conditions are fulfilled:

Condition I: Existence of similar morphemes:

- (i) Genetically related languages A and B always share a sufficient number of similar morphemes<sup>10</sup>.

If the two languages reveal a sufficient number of similar morphemes, one can assume that this similarity is not accidental, and it can probably be interpreted as evidence of genetic relation. Similar morphemes can be either lexical or grammatical, but the existence of similar lexical morphemes seems to be obligatory: there is no generally accepted language family for which languages do not share similar lexical morphemes, while for several well-established language families of Southeast Asia (Kadai, Vietic, Lolo-Burmese) no grammatical morphemes are known so far (Peiros 1998). If similar morphemes are not found, we do not have data for a further discussion of genetic relationship of these languages.

Condition II: Genetic reasons for the similarities between morphemes:

- (ii) Sufficient number of similar morphemes in languages A and B belong to the core lexicon.

Similar morphemes can be found in all parts of the lexicon, but if the languages are genetically related, they always share morphemes from the core lexicon. We will develop and discuss this notion later (§19). Here it is enough to say that core lexicon includes words with simple universal meanings, which are less open to borrowings than other parts of a language lexicon. There is no doubt that words from the core lexicon can also be borrowed but the likelihood of borrowing here is usually lower. To the best of our knowledge, all known related languages always share words from the core lexicon. Thus one can conclude that if two languages share not only similar morphemes but morphemes which also belong to the core lexicon, it is more probable that these languages are genetically related.

Similarities between morphemes can be due to various reasons: common origin, borrowing, chance resemblance, and so on. To demonstrate the genetic

---

<sup>10</sup> For other views see, for example, Guy 1981, Nichols 1996.

nature of these similarities, we need a system of phonological correspondences between the languages:

Condition III: Existence of systemic phonological correspondences:

- (iii) The phonological systems of A and B are connected by systemic phonological correspondences<sup>11</sup> with the element of one system corresponding to certain elements (one, several or none) in another.
- (iv) The systemic phonological correspondences mentioned in (iii) are true for lexical similarities discussed in (i) and (ii).

These conditions are sufficient to provide us with formal criteria to judge if there is enough evidence to accept that two languages are genetically related and (due to transitivity of the notion<sup>12</sup>) that all languages related by them belong to the same linguistic family. It is important to mention that there is no additional requirement for grammatical similarities. However, where there are such similarities, they can provide an additional and often crucial support for a genetic claim.

The conditions discussed require that a set of systematic phonological correspondences be established between all genetically related languages.

However, to be able to establish such correspondences we need a certain level of understanding of the languages' relations. It would not be wise, for example, to try to obtain such correspondences investigating simultaneously English, Dutch, Russian and Polish. The more natural procedure would be at first to study separately Germanic and Slavic languages and then to compare the results. To do this we need, however, to be able to identify these two groups at least at the level of intuition. The whole process thus can be represented as consisting of four consecutive stages:

- (1) a language family is identified using various heuristic procedures;
- (2) a hypothesis about its classification is suggested;

---

<sup>11</sup> A phonological correspondence is a systemic one if it brings together reflexes of a particular proto phoneme. By the definition, a correspondence based on borrowings cannot be a systemic. A systematic correspondence can be either regular (e.g. found in many examples) or not.

<sup>12</sup> If language A is genetically related to language B and language B is genetically related to language C, then A is always genetically related to C.

(3) the formal method of comparative reconstruction is applied to the languages intuitively included in the family;

(4) a classification based on the results of the reconstruction is suggested.

The differences between stages 1/3 and 2/4 affect the reliability of the results and it is quite possible that some of the decisions made at the intuitive stages 1 and 2 will be rejected by the strict procedures of the analysis conducted at the stages 3 and 4. At the same time, one cannot start an investigation just from the third stage ignoring all intuitively based decisions.

Let us now limit our discussion to stage (4) which means that we are dealing with a well-established language family.

§13. A language family is formed by processes started in the past but with results observed in the present time. Studying these processes, linguists try:

- to reconstruct the common proto-language of the family and thus to explain the origins of structures of recorded language;
- to suggest a model of family evolution from its proto-language into historically attested languages.

The reconstruction of the common proto-language is conducted in several steps. It starts with the investigation of a group of recorded sister-languages and with a reconstruction of their proto-language. At the next stage of research, the same procedure is repeated, but instead of the recorded languages, we work with their reconstructed proto-language which is compared with its sister-languages, either recorded or reconstructed. Their ancestor is reconstructed which in turn, will be later used for more ancient reconstruction. The theoretical requirement is, that a daughter-language should never be used instead of its reconstructed ancestor: as soon as the reconstruction is completed, we should act as if more recent stages of its development do not exist at all (= what is not included in the reconstruction did not exist in the proto-language).<sup>13</sup>

The procedure of reconstruction thus goes in the opposite direction to the real process of languages development: starting with present recorded languages or

---

<sup>13</sup> We are free, however, to modify our reconstructions to include previously unaccounted features of daughter-languages.

relatively recent recordings, it moves back in history discovering at each step more and more ancient ancestors of these languages. This retrospective approach is the only justifiable approach in comparative linguistics.

A creation of an evolutionary model follows the same pattern:

- we move in the direction opposite to the real process: from the present to the past;
- at any taxonomic stage *S*, we identify sister-languages *A*, *A'*, *A''*, which are all daughter-languages of *L*;
- at the chronologically preceding and thus more ancient taxonomic stage *S*<sub>1</sub>, a search for languages specifically related to *L* (*L'*, *L''*, etc.) is conducted and their common proto-language *T* is identified. This is done with no reference to the situation at stage *S* (*A*, *A'* and *A''* or to daughter-languages of *L'* or *L''*).

§14. At any period of its evolution, a language family is characterized by various degrees of differences accumulated in its languages. Therefore we need to differentiate two aspects of family evolution: a chronological aspect (languages exist in time and in any given period of time where they are at a particular stage of their development) and a 'divergent' aspect (the increase of differences between the languages). Ideally, any model of a family evolution should:

- be chronologically correlated, telling us about the relative chronology of languages of the same or different taxonomic levels ("Did the split of Proto Germanic take place much later than the split of Celtic?"). Chronological information, at least a relative one, seems to be an essential part of any model dealing with processes, including language development.

- provide us with information about diversity among the related languages which range from the cases where it is hard to decide if we are dealing with dialects of the same language or with two closely related languages to the cases where very complex research should be undertaken to demonstrate the very fact of the relationship. In the former case, the systems are quite similar while in the latter, only obscure traces of the similarities can be identified. Dealing with various groups of related languages, linguists always want to know how different the languages are within a group in comparison to another, usually better known group: "Are Eastern Slavonic languages closer to each other than the Southern Slavonic ones?", "Are



Zhuang-Thai languages less genetically diverse (closer to each other) than Slavonic ones?”.

These considerations suggest that an ideal model of a family evolution should represent:

- the structure of the family (= its genetic tree);
- the family's internal diversity;
- chronological stratification of the family's evolution.

The problems of chronological stratification belong to the theory of glottochronology and are not discussed here (see Starostin 1989). This leaves us with a partial model dealing only with branching and degrees of similarities.

§15. The development of a family structure, i.e. its branching, is connected to the notion of language continuity and thus the corresponding component of the model has to be based on pure genetic data.

Internal diversity of a family is determined by the accumulated amount of both internally and externally triggered changes. So evaluating these degrees we have two options:

- (i) either to use the whole range of data, including similar loans, regional features, and other non-genetic features;
- (ii) or to concentrate only on genetically caused features preserved by the languages from their common ancestors.

The second option (orientated towards genetically caused similarities) opens a possibility to use the same type of data as for modelling branching.

§16. The process of evolution of a language family based on the split of an original single language into its descendants is a universal one and can be observed in all language families regardless of the internal organisation of languages, the specific features of speech communities and other circumstances. Therefore it should be possible to suggest a universal model of evolution applicable to any language family of the world.

For example, we should be able to analyse in the same way the Vietic family of Austroasiatic (known for its lack of morphology and intensive contacts with the

other languages of the Southeast Asian linguistic area), the Paman family (a typical Australian family both in its structural features and cultural and sociolinguistic characteristics of its speakers), and the Slavonic languages of Indo-European. Obtaining evolutionary models of these three families should enable us to compare the relation between Vietnamese and Arem of Vietic to that of Yir-Yoront and Jabugay of Paman or Bulgarian and Russian of Slavonic. It should also enable us to answer questions such as: 'which of these pairs represents sister-languages?', 'which pairs are more diverse?'.

By studying a family's evolution, we model a real historical process. Our model therefore has to meet at least the following conditions:

I. As it deals with genetic relations between languages, it has to be based on pure genetic considerations and thus:

- criteria used in the classification have to be of pure genetic nature;
- these genetic criteria have to be made explicit.

II. The procedure used should be universal and applicable to any language family, regardless of their typology and other characteristics; so:

- the procedures must use features found in all human languages;
- the treatment of these features should be identical for any language family.

Therefore the evolutionary trees are to be comparable across language families.

III. An evolutionary tree generated by the model represents real historical relations between languages and thus:

- it should provide information about:
  - (a) the structure of the family (= its genetic tree)
  - (b) the internal diversity of the family;
- a procedure of classification has to be automatic and free of any type of personal preferences of scholars involved:
  - (a) two scholars working independently should produce the same tree;
  - (b) the choice of parameters for classification should be based on objective rather than subjective criteria;

- (c) the results of a classification should be open to a formal procedure of evaluation.

IV. The method of classification should be reasonably simple.

§17. Let us discuss now how to meet these requirements and to create an evolutionary model.

The genetically caused similarities between languages A and A' or B and B' of Figure II are due to the retention of features developed in periods I and II, while their differences are the result of their independent development in period III. In the same way the genetically caused similarities between A and B or A' and B can only develop during period I, while their differences have appeared in periods II or III.

Theoretically one can expect to find that:

- the total amount of genetically caused similarities between A and A' should always be higher than between A and B or A' and B', as the existence of A and A' as a single entity was longer and thus more features were adopted;
- some of the genetically caused similarities between A and B or A and B' should be more or less identical: having retained from *T*, the common ancestor of the family and thus having a good chance of being retained with the descendants;
- the amount of genetically caused similarities between any pair of languages reflects the level of their relationship and thus their position in the genetic tree.

Here we would expect to hear the following argument. "Imagine, that language A, in the process of its internal development (period III), has lost all features inherited from L. In such a case the amount of its similarities with A' would not be different from that with B and thus it would be impossible to demonstrate its specific relation with A'. That is why the above mentioned considerations are not convincing". Such an argument is one of the numerous linguistic myths besieging lexicostatistics. There are no proven records of such developments, i.e. information about well-defined language groups where languages have no traces of a specific common origin. The standard procedure of

comparative linguistics would not be able to detect them and justify that A and A' should be kept together. Without common features, no substantial comparative claim about specific relations can be made<sup>14</sup>.

Comparing the systems of  $T$  and its daughter-language  $L$  one finds:

- (i) R(etentions) - features retained in  $L$  from  $T$  (result of drifts);
- (ii) A(dditions) - features added to the system of  $L$  in the process of its crystallization or integrated development (result of additions);
- (iii) S(ubtractions) - features of  $T$  not retained in  $L$  (result of losses).

Let  $\Sigma(L)$  be the set of structural features of  $L$ . By definition  $\Sigma(L)$  includes different features of  $L$ : its phonemes, morphemes, grammatical and syntactic rules and so on. As we can separate  $L$  from all other languages, including its direct ancestor  $T$ ,  $\Sigma(L)$  is always different from  $\Sigma$  any other language.

Let  $R(L, T)$  be the set of features retained in  $L$  from its ancestor  $T$ . This  $R(L, T)$  is always smaller than  $\Sigma(L)$ , as it also includes  $A(L)$  - a set of specific additions to  $L$  not found in its ancestor. Therefore  $\Sigma(L) = R(L, T) + A(L)$ .

If  $L$  and  $L'$  are sister-languages, they retained features of their common ancestor  $T$  and those features are the only source of genetically caused similarities found in the two languages.  $L$  and  $L'$  have developed separately and thus their losses of the original system cannot be identical. Some of them could coincide, but the total set of losses would never be the same. As two languages  $L$  and  $L'$  can never develop in exactly the same way, each has language-specific changes and  $R(L, T)$  is always different from  $R(L', T)$ .

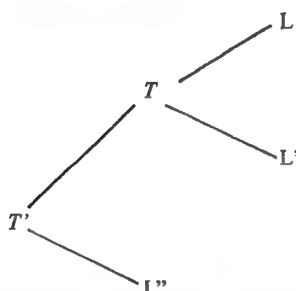
Let  $D(A, B)$  - we will call it '**genetic distance** between A and B' - be a set of genetically identical features retained in A and B. For  $L$  and its direct ancestor  $T$   $D(L, T)$  is identical to  $R(L, T)$ . For two sister-languages  $L$  and  $L'$ , their  $D(L, L')$  is an intersection of two sets:  $R(L, T) \cap R(L', T)$ .

For any  $L$ , its  $D(L, L')$  is always smaller than its  $D(L, T)$ , but is always greater than  $D(L, B)$ , where  $B$  is not  $L$ 's direct ancestor or a sister-language. This claim is based on the following arguments. For the family represented in Figure III we will always find more similarities between  $L$  and  $L'$  than between  $L$  and  $L''$  or  $L'$  and  $L''$ . Shared features of  $L$  and  $L'$  are inherited from  $T$ , while shared features, say, between  $L$  and  $L''$  are inherited from  $T'$ .  $L'$  and  $T$  are sister-languages and their

<sup>14</sup> A discussion of 'retention rates' is given below (§ 26).

sets of shared features are always smaller than the sets of features inherited by each of them from their common ancestor  $T'$ .  $L$  and  $L'$  are daughter-languages of  $T$ , which is the source of all their common features. This source, as we have already discussed, is formed by retentions from the previous stage as well as by specific additions not known in its sister languages:  $R + A$ . This unique set is the original source for  $L$  and  $L'$ . It is absolutely impossible that one of these languages would lose all the features specific to its ancestor adopting at the same time some features from the ancestor sister-language  $L''$ . Therefore  $D(L, L')$  is always smaller than  $R(L, T)$  or  $R(L, T')$ . As  $D(L, L'')$  is in turn always smaller than  $R(L, T')$  it should be also smaller than  $D(L, L')$ .

Figure III  
(same as Figure I)



If we are now able to suggest a procedure to identify and compare these genetic distances we should be able to answer the following questions:

- (i) what languages are directly related?
- (ii) which of them are sister-languages?
- (iii) what language is the common ancestor of the family?

This procedure will provide us with a formal tool to create genetic trees and to represent the internal diversity of the family, thus achieving (apart from getting absolute datings) objectives of the evolutionary model.

§18. To evaluate genetic distances we can either use all the information available for these languages or choose certain sets of parameters which would allow us to achieve the same goal in a more economic way. The second approach seems to be preferable, but raises two questions: 'how to select the parameters?' and 'how to use them to obtain a classification?'.

Let us first discuss how to select the parameters. The evolution of a language family is caused by various changes whose accumulated results led to the disintegration of a former single language. Therefore it seems logical to choose parameters related to such changes.

We have discussed already the difference between internal and external changes. Of these only internally caused changes can be used to investigate genetic relations between languages. External changes are caused by foreign influence and thus they cannot indicate pure genetic relations. Rather they can be used in classifications whose goal is to identify linguistic areas ('Balkan languages', 'Southeast Asian languages') or groups of languages under a common influence ('languages of Sino-centric world'), but not a genetic relation. Therefore for each change (or its result) used in any genetic model we need to provide evidence of its internal nature. Such evidence presented explicitly must always be included in the data.

There are two main ways to select diagnostic changes:

- to create a fixed list of them to be applied to various languages; or
- to suggest a formal procedure of their selection, which would provide us with different lists depending on the family to be investigated.

As it is highly unlikely that a change can be found in all human languages, it seems preferable to deal with a selection procedure, rather than with a certain list, knowing that such a list is by no means universal.

The selection of diagnostic parameters determines the whole model, and should reflect the requirements outlined above:

- selected parameters are to be found in all human languages;
- the parameters are to be either internally caused changes or the reflexes of such changes;

- selection of parameters should always be conducted automatically, i.e.; not affected by the personal preferences of a researcher: 'I think that this change is revealing, but I cannot prove this'.

This leads us to the question of 'what kind of parameters can be chosen for the model?' We have decided already that the parameters chosen have to be either internally caused changes or their reflexes. However it seems also important to use only free and not triggered changes. The reason for such a limitation follows.

Imagine that we are dealing with a language *T* which distinguishes ten noun cases. The morphs for three of them - Accusative, Genitive and Dative - differ only by final stops: *at*, *ap* and *ak*. In the history of its daughter-language *L* final stops were lost which triggered the loss of distinctions between these three cases and required restructuring of the whole case system. In *L'*, another daughter-language of *T*, final stops are preserved and the original case system remains unchanged. If we now choose the case markers to be our parameters, we would most probably come to a wrong conclusion. In this example, only the phonological change can be selected as a parameter to be used in the model, provided that it has not been triggered by other changes.

§19. In any evolutionary model, we are dealing with genetically related languages, which means that all of them are daughter-languages of the family proto-language *T*. This means that there is always an uninterrupted language continuity between *T* and its daughter-languages. We have discussed already (§4) language continuity defined with the help of core meanings (over the whole period of language continuity, the core information is conveyed mainly with the help of linguistic means of the same origin).

"As is well known, natural language has only two major types of expressive means—lexical and non-lexical—to encode the information a sentence carries. Lexical means are simply words. The set of lexical means used in a sentence is the list of all wordforms, or lexeme occurrences, that constitute it.

Non-lexical means are of three varieties:

- linear order of wordforms
- prosody (intonation contours, pauses, phrase and sentence stress)
- inflections (i.e. morphological categories).

Of these three varieties, word order is the most important and the most universal, being necessarily present in every language and in every sentence. It is imposed by the physiologically conditioned linearity of human speech. Inflections, on the other hand, are the least important ... because they are the least universal linguistic means: some languages lack them entirely; in all

languages there are uninflected words, which, however, are syntactically linked to other words in a sentence. Prosody occupies an intermediate position.

There are no other types of linguistic expressive means.

Both lexical and non-lexical expressive means of a language can be used in one of the following two ways:

- Either in a SEMANTIC capacity, i.e., to convey meaning immediately...
- Or in a SYNTACTIC capacity, i.e. to mark relations between linguistic entities..." (Mel'chuk 1988, p.19-20).

Usage of lexical means in a syntactic capacity or non-lexical means in a semantic capacity is always language-specific. Therefore, if we are planning to use expressive means of languages to build a universally applicable model of a family's evolution, we can employ only lexical means in a semantic capacity and/or linear order of word forms used in a syntactic capacity.

Any model of a family's evolution establishes continuity between languages which has been defined with the help of core meanings, so it seems natural to build a model also orientated towards semantics, or more precisely to the lexical means of languages used in their semantic capacity. Otherwise we will adopt a logically inconsistent approach with language continuity defined with the help of core meanings, while other features are used to identify this continuity<sup>15</sup>.

Core meanings are found in any human language where they are normally represented by various lexical means of this language, words or more complex structures. These lexical means form the **core lexicon** of the language. As there is no apparent reason for these core meanings to be dropped from the language, we can assume that the corresponding words have a better chance of being preserved than words from other parts of the lexicon. At the same time, it is well known that words from the core lexicon can also be borrowed or substituted due to internal development. In all such cases, the corresponding meanings nevertheless remain in the language.

The core lexicon of language A is a set of words or more complex lexical structures whose meanings are the most precise equivalents of the core meanings. The core lexicon of language B is another set of words which represent the same set of meanings. These two core lexicons are not identical but we can expect that for every word from A, there is one or several words in B whose meanings are

---

<sup>15</sup> It is extremely important to mention here, that we do not claim that other features (like genetically caused irregularities of verbs in Germanic languages) do not indicate specific genetic relations. The claim is that a universal model should be built on an analysis of core meanings because of their universal nature.



identical and represent a particular core meaning [. These words are [-synonyms. As the core meanings are supposed to be universal, it is theoretically possible to identify the core lexicon in any human language. All such core lexicons would include forms with comparable meanings.

§20. We have already formulated the three basic logical assumptions of the evolutionary model:

(1) any two genetically related languages have common features which are retained in these languages from their common ancestor<sup>16</sup>;

(2) these common features can be found both in the lexicon and grammar (in the wider sense) of these languages. There is, however, no well established language family without common lexical similarities;

(3) there is always a correlation between the position of languages in a classification and the amount of genetically caused similarities between them. If two languages A and B are grouped together, they always retain more common features than with any other, more distantly related language, C. If A, B and C are equally distant in a classification, no pair of them reveals a significantly higher amount of common inherited features;

To these, we can add now a fourth assumption:

(4) the same correlation is observed if we compare the core lexicons of languages: if languages A and B are closer to each other than C, more words of common origin are found in their core lexicons, and fewer are shared by A and C or B and C.

Now we can formulate two basic postulates of lexicostatistics:

1. Core lexicons can be used to identify genetic distances between related languages and to model the evolution of the whole family.
2. It is possible to do this with the help of a limited sample chosen from the core lexicons of related languages.

§21. We have already discussed how a language's core lexicon is formed by words of this language which represent the core meanings. These meanings are,

---

<sup>16</sup> An absence of such similarities indicates the absence of transparent relationship which, however, does not rule out that the languages may be distantly related.

by definition, universal, while the corresponding words are language-specific. It follows from the proposed assumptions that it should be possible to compile a list of core meanings suitable for an evolutionary model. Such a list, let us call it ' $\sigma$ -list', should have the following features:

1. it is formed by core meanings which, by definition, are usually represented in all human languages regardless of habitat, economic activities or other characteristics of their speakers;
2. it includes easily identifiable meanings which are normally associated with simple words in languages. From this point of view,  $\sigma$ -list should not include such words as 'life' or 'anger', as for many languages it would be quite difficult to identify corresponding words;
3. the chosen meanings should be free from taboos and other cultural restrictions on their possible usage. Thus we cannot include in the list such core meanings as 'to give birth' or 'excrements';
4. meanings included in  $\sigma$ -list should be historically independent from each other and a change in one meaning should not trigger changes in other meanings from the same list. From this point of view, the list is not supposed to include meanings from the same semantic field, like 'eye', 'pupil of the eye', 'white of the eye', and others, as a change in one of them can easily cause a domino-like effect for changes among the others.

If these conditions are met, it is not very important how many meanings are included in the  $\sigma$ -list. However, a very short  $\sigma$ -list (say 20 items) is not a good option, as the impact of each entry becomes too great. On the other hand, an analysis of a very long list (say 1,000 items) can be excessively time consuming and thus is not convenient. It is important to mention, however, that differences in  $\sigma$ -list can significantly affect the evolutionary tree obtained, so that one cannot compare results obtained with the help of a 200-item list with the results obtained with the help of 100-item or 35-item lists. Even differences of lists in 10 - 15 meanings may make the trees incompatible.

Figure IV  
Standard  $\alpha$ -list

- |  |  |   |
|--|--|---|
| 1. all (as in 'all of the stones')         | 38. head                                 | 72. seed (of a fruit)                             |
| 2. ashes (cold ashes left after fire)      | 39. hear (as sounds)                     | 73. sit (as in 'he is sitting here')              |
| 3. bark (of a tree)                        | 40. heart (as body part)                 | 74. skin (of human being)                         |
| 4. belly (the outside part of human belly) | 41. horn (cow's horn)                    | 75. sleep (as in 'he is sleeping')                |
| 5. big (big in size, as opposed to small)  | 42. I                                    | 76. small (in size, opposite to big)              |
| 6. bird                                    | 43. kill (as a person)                   | 77. smoke (of a fire)                             |
| 7. bite (bite as in eating)                | 44. knee                                 | 78. say or speak (as in 'he is saying something') |
| 8. black (black colour)                    | 45. know (as in 'I know his')            | 79. stand (as in 'he is standing there')          |
| 9. blood                                   | 46. leaf                                 | 80. star  |
| 10. bone                                   | 47. lie (as 'the boy lies on the floor') | 81. stone   |
| 11. breast (female breast)                 | 48. liver                                | 82. sun   |
| 12. burn (vt., to burn sticks)             | 49. long (rope)                          | 83. swim (as in 'he is swimming there')           |
| 13. claw or fingernail                     | 50. louse (hair)                         | 84. tail (of a dog)                               |
| 14. cloud                                  | 51. man (as opposite to woman)           | 85. that (far from here)                          |
| 15. cold (as water)                        | 52. many (as in 'many stones')           | 86. this (close to here)                          |
| 16. come                                   | 53. meat (as in 'to cook meat')          | 87. thou (2 <sup>nd</sup> Sg.)                    |
| 17. die (of a person)                      | 54. moon                                 | 88. tongue (of a human being)                     |
| 18. dog                                    | 55. mountain                             | 89. tooth   |
| 19. drink (as water)                       | 56. mouth (of a person)                  | 90. tree  |
| 20. dry (as clothes)                       | 57. name (of a person)                   | 91. two   |
| 21. ear                                    | 58. neck (as opposite to throat)         | 92. walk or go (as in 'we will walk there/here')  |
| 22. earth (as in 'a shovel of earth')      | 59. new (as in 'new clothes')            | 93. warm (as water)                               |
| 23. eat                                    | 60. night (as opposite to day)           | 94. water (fresh water)                           |
| 24. egg                                    | 61. nose                                 | 95. we (Pl. excl.)                                |
| 25. eye                                    | 62. not (as in 'not new')                | 96. what?   |
| 26. fat (n., as in 'meat and fat')         | 63. one                                  | 97. white (white colour)                          |
| 27. feather                                | 64. person (= human being)               | 98. who?  |
| 28. fire                                   | 65. rain                                 | 99. woman (as opposite to man)                    |
| 29. fish (n.)                              | 66. red (colour)                         | 100. yellow (yellow colour)                       |
| 30. fly (as a bird)                        | 67. road or path                         |   |
| 31. foot                                   | 68. root (of a tree)                     |   |
| 32. full (as a basket)                     | 69. round (object)                       |   |
| 33. give                                   | 70. sand                                 |   |
| 34. good                                   | 71. see (as in 'I can see him')          |   |
| 35. green (colour)                         |  |   |
| 36. hair (of head)                         |  |   |
| 37. hand                                   |  |   |

There are many ways to compile a  $\sigma$ -list, but we prefer to use the standard one which includes 100 meanings suggested by Swadesh, despite the fact that the list is not ideal as:

- some of its meanings are not universal: for example, until recently the meaning 'horn' was not known in Australian Aboriginal languages;

- not all of its meanings are completely independent: as, for example, in the case of 'bark' and 'skin';
- not always are the meanings free from cultural impact: it is possible, for example, the meaning 'person' which often has connotations like 'us, true human beings'.

Nevertheless it seems convenient to deal with the standard list, as all other  $\sigma$ -lists always have problems of their own. One advantage of the standard list is that it has already been used for many languages, which helps in collecting data and comparing results.

§22. The actual lexicostatistical procedure starts with the completion of the **diagnostic lists** ( $\sigma$ -lists) of all the languages under investigation.

A  $\sigma$ -list of language L is formed by words or more complex lexical units, whose meanings are the most precise unmarked equivalents of corresponding  $\sigma$ -meanings. It represents this language as it is used by a certain well-defined group of people in a particular period of time (it is dialectally and chronologically specific). Therefore it always has to include only forms taken from a certain dialect rather than from several different dialects or even closely related languages. It also has to represent a particular chronological period and not include forms used in different periods of the language's history.

A word included in a  $\sigma$ -list has to be the most precise translation of the  $\sigma$ -meanings into the language and thus be:

- (i) the most common representation of a  $\sigma$ -meaning in the language. We have to include only the most widely used word, as for example, the first word given for an English-L dictionary entry. Such dictionaries are often the best source of data for compiling a  $\sigma$ -list if a judgment of native speakers is not available.
- (ii) taken from the unmarked, neutral style or register of the language, as otherwise the word would not be the most precise representation of a  $\sigma$ -meaning. If for example, word *traj* means 'tail', but is used only by hunters, while other speakers use another word *kan*, only the latter is to be included in the  $\sigma$ -list.

Taboos raise a specific theoretical problem. In taboo situations a particular word can temporarily not be used and another word used instead. Over a certain period of time, the original word would regain its unlimited role. However, to observe a taboo, a speaker is supposed to know both words (tabooed and its substitution), but to use only the substitution. We may include both forms in a

$\sigma$ -list, but it seems more logical to deal only with the original word, treating its taboo substitution as a less general, marked form with no reason to be included in a  $\sigma$ -list. It is worthwhile to mention, however, that clear taboo substitutions have not been encountered in our work with various language families, including the Australian Aboriginal languages, known for their taboo replacements (Dixon 1980:28; Alpher & Nash, forthcoming). All languages' dictionaries usually provide us mainly with main words, while taboo substitutions, if included, are always specifically marked.

(iii) The process of the completion of a  $\sigma$ -list should not be affected by any historical considerations, and every  $\sigma$ -list should represent the typical usage of the period chosen. It means that we should not include in a  $\sigma$ -list forms which have good etymologies, but are not central from the point of their usage. For example, despite the fact that archaic Russian *oko* 'eye' has a good Indo-European etymology, the Russian  $\sigma$ -list includes only the modern form *glaz*.

(iv) Sometimes a language does not have one single word whose meaning corresponds precisely to an  $\sigma$ -meaning. Two different situations are represented here:

A language does not have a word with a  $\sigma$ -meaning, instead a word with a broader meaning is used. Russian, for example does not distinguish between the meanings 'leg' and 'foot'. As the corresponding word - *noga* - is the most precise representation of the  $\sigma$ -meaning FOOT, this word is included in the Russian  $\sigma$ -list.

Nyawaygi, an Australian Aboriginal language, does not have a single word which represents the  $\sigma$ -meaning MOON; instead it has two different words *palanu* 'new moon' and *ilkan* 'full moon' (Dixon 1983). Both words should be included in the  $\sigma$ -list of Nyawaygi, thus achieving the aim to have the corresponding  $\sigma$ -meaning be represented completely.

If these requirements for the selection of candidates to a  $\sigma$ -list are met, these properly chosen words are called (proper) **list-members**.

$\sigma$ -lists of various languages have different sets of forms, but they always include forms which represent only one hundred core meanings. This fact is used when we compile **lexicostatistical tables**. A lexicostatistical table includes all  $\sigma$ -synonyms (one or several words) found in all the languages analysed. Table HAND, for example, includes forms of languages which are synonymic and mean 'hand'. This meaning can be represented by a word which means only 'hand', by a word which means 'hand / arm', by a simple word, a compound or even by several words. Each lexicostatistical table has two columns: one lists the languages under

investigation, while another lists the forms of these. Each line of a table thus connects a language with its form. As [-list is formed by one hundred meanings, the whole database of the analysis includes one hundred lexicostatistical tables.

The lexicostatistical table 'STONE' for several Central Pacific languages<sup>17</sup> of the Austronesian family is:

Languages	Forms
East Fijian	βatu
West Fijian	βaču
Rotuman	hofu
Mele-Fila	fatu
Tahitian	?oofaʔi
Rapanui	ma'ea
Nukuoro	hadu
Maori	koohatu, poohatu
Samoan	maʔa
Tongan	maka
Hawaiian	poohaku

§23. The next step of the procedure is the etymological identification of forms presented in individual lexicostatistical tables. For each word J of language L included in table T, we have to give Yes /No answers to two questions:

- is β a borrowing?
- does β have the same origin as form γ of L' in the same table?

Two comments are needed here:

(i) the etymological identification of the forms given in table T is not a complete etymological analysis. The aim of the latter is to find all cognates preserved in the languages, including those which changed their meanings: 'meat' => 'bird' or 'eye' => 'face'. Such forms with different meanings can sometimes be included in several different lexicostatistical tables or be left outside the investigation. Etymological identification in lexicostatistics deals only with data included in one particular table T and discusses only one question: if forms β and γ

<sup>17</sup> Lexical data is taken from Tryon 1995 and several language dictionaries.

of languages A and B included in T are cognates, or in other words, whether both J and K represent independent uninterrupted developments of the same *T* of their common ancestor. From this point of view, the fact that the word 'water' of language A is a cognate to the word 'rain' of language B and to the word 'cloud' of language C is irrelevant, as these words are included in three different lexicostatistical tables.

(ii) in the procedure, we analyze only lexical morphemes of the words included in table T. Affixes are not analyzed. This restriction follows from the orientation of the whole model towards lexical rather than grammatical means of a language.

Etymological identification is supposed to be based on detailed knowledge of the historical phonology and etymology of the family we investigate. Without it, one cannot conduct reliable identification of borrowings and cognates and thus full lexicostatistical analysis is not applicable<sup>18</sup>.

After etymological identification is conducted, table T is given a the third column: information about the origins of all lexical morphemes included in it. This information is represented in numerical form:

- negative numbers for loans
- positive numbers for original words.

Words of the same origin have identical (positive or negative) numbers.

The etymologized lexicostatistical table 'STONE' for languages given above is<sup>19</sup>:

Languages	Forms	Etymological information
East Fijian	βatu	1
West Fijian	βaču	1
Rotuman	hɔfu	2
Mele-Fila	fatu	1
Tahitian	ʔoofaʔi	3
Rapanui	maʔea	4

<sup>18</sup> One can, however, use 'Preliminary lexicostatistics' as a heuristic method (see Peiros, in this volume).

<sup>19</sup> Etymological information is taken from Biggs MS.

Nukuoro	hadu	1
Maori	koohatu, poohatu	1
Samoan	maʔa	4
Tongan	maka	4
Hawaiian	poohaku	1

The forms identified with the number 1 are reflexes of Proto Austronesian *\*batu*; the forms identified with the number 4 are reflexes of Proto Polynesian *\*maka*; the forms identified with the numbers 2 or 3 are isolated in the table.

The same table for several Mon-Khmer languages is:

Languages	Forms	Etymological information
Katu	dəl	1
Bruu	təməw.L	2
Kui	təmau.L	2
Pakoh	bul	3
Wa	si.mauʔ.B	2
Lawa	səmoʔ	2
De'ang	mau	2
Plang	kaʔ.4 muʔ.2	2
U	mo.2	2
Khmu	klà:ŋ	-4
Ksinmul	ʔəlɛŋ	-4

The forms identified with the number 2 are reflexes of Proto Mon-Khmer *\*Cəmauʔ* > Proto Katu *\*[t/d]əmhaw* (Peiros 1996 N 475) and Proto Palung-Wa *\*səmauʔ* (Peiros MS). The forms identified with -4 are loans from a Sino-Tibetan source: Proto Sino-Tibetan *\*Lə:ŋ* / *\*Lə:k* (Peiros & Starostin 1995, 3:69)

Further analysis is based on these etymologically supported numbers and not on the actual forms given in the tables. We do not need to know any more that forms in Table STONE are *stone* and *Stein* as in English and German or *maʔa* and



*maka* as in Samoan and Tongan. All we need to know is that in both cases we are dealing with genetically identical forms.

§24. The **lexicostatistical data-base** for a family is formed by one hundred lexicostatistical tables which include:

- (i) forms from all the languages under investigation;
- (ii) the etymological identification of each form included in a table.

The next step of the procedure is a statistical analysis of this data-base and the completion of a lexicostatistical matrix of a family. The basic idea of this procedure is the following: for each pair of languages, we calculate the percent of etymologically identical words and include them in a matrix.

Borrowings and situations with more than one form represented in a language require special discussion.

When dealing with borrowings, we can choose one of the following strategies:

(i) we can treat borrowings in the same way as original words. With such an approach, borrowings from the same source are to be analyzed as genetically caused similarities. This would affect the amount of similarities between the languages which in turn would affect their position in the evolutionary model. As we have already decided that this model represents only genetic relations of languages and not their contact, cultural, typological or other relations, such an approach cannot be accepted;

(ii) we can agree to treat borrowings (regardless of the fact that they can be of identical origin) as a lack of genetic identity. This would lead us to the same problem as in (i), i.e. an effect of borrowings on a genetic classification;

(iii) borrowing is not a genetic process and its results - various loans, simply substitute the original words which are not any more available for our investigation. Therefore, it seems logical to treat loans as lack of information, rather than to include them in an analysis of genetic development.

Forms of two languages in a table can be related in three different ways:

(a) L	L'	(b) L	L'	(c) L	L'
$\alpha \Leftrightarrow \gamma$		$\alpha \Leftrightarrow \gamma$		$\alpha \Leftrightarrow \gamma$	
		$\beta \parallel \delta$		$\beta \Leftrightarrow \delta$	

If we treat (c) as representing two separate cases of genetic identity:  $\alpha \Leftrightarrow \gamma$  and  $\beta \Leftrightarrow \delta$ , then for (b) we have to talk about one case of identity ( $\alpha \Leftrightarrow \gamma$ ) and one case of lack of identity ( $\beta \parallel \delta$ ). It means that statistically (c) would be treated as 2 identities, (b) - as 0 (one identity + one lack of identity = 0) and (a) as 1 (one identity). Such an approach does not seem to be the best one, and for every pair of languages, we count only identity regardless of the fact that in reality we have more than one pair of etymologically identical morphemes. In the situation:

$$\begin{array}{ccc} L & L' & L'' \\ \alpha \Leftrightarrow \gamma & & \\ \beta \Leftrightarrow \delta & & \\ & \varepsilon \Leftrightarrow \zeta & \end{array}$$

we will accept only one countable identity for L and L' and one for L' and L''.

§25. There are three types of relations between related languages:

1. Specific:

1.1 the relation between an ancestor and its daughter-languages;

1.2 the relation between sister-languages (the languages with the same direct ancestor)

2. Non-specific: the relation between other types of related languages.

It follows from the above discussion, that in a lexicostatistical matrix we expect to find that specifically related languages are marked by a higher percent of common shared words across the whole lexicostatistical data-base.

Let us examine the matrix in Figure V.

Figure V: A sample lexicostatistical matrix

	A	B	C	D	E	F	J
A		75%	40%	42%	40%	42%	38%
B	75%	-	40%	42%	40%	42%	38%
C	40%	40%	-	60%	50%	52%	50%
D	42%	42%	60%	-	52%	48%	50%
E	40%	40%	50%	52%	-	70%	68%
F	42%	42%	50%	48%	70%	-	72%
J	38%	38%	50%	50%	68%	72%	-

We can identify three groups of specifically related languages in this matrix: A/B, C/D and E/F/J. The percentages shared by languages of each of these groups are higher than for the languages across the groups:

	A	B	C	D	E	F	J
A		75%	40%	42%	40%	42%	38%
B	75%	-	40%	42%	40%	42%	38%
C	40%	40%	-	60%	50%	52%	50%
D	42%	42%	60%	-	52%	48%	50%
E	40%	40%	50%	52%	-	70%	68%
F	42%	42%	50%	48%	70%	-	72%
J	38%	38%	50%	50%	68%	72%	-

If so, the languages of every group have to have a common ancestor: L for A and B, L' for C and D, and L'' for E, F and J. These ancestor-languages existed earlier than the period represented by matrix I. We can represent the relations between the three ancestor-languages in matrix II:

	L	L'	L''
L (A/B)	-	41%	40%
L' (C/D)	41%	-	50%
L'' (E/F/J)	40%	50%	-

Higher percentages between L' and L'' suggest that these two languages are specifically related and at an earlier stage of family's evolution they were represented by a single common ancestor.

Comparing core lexicons of language L and its ancestor T, we can find the same picture as we have already discussed:

- (i) words retained in L from T;
- (ii) words known in L, but not in T. The words have appeared in L after it separated from T due to various additions to L's lexicon.

The balance between groups (i) and (ii) reflects the level of similarities between these languages: more closely related languages always have more common words. This theoretical suggestion is based on our experience in comparative linguistics, as this balance can be observed for any pair of well studied languages. Therefore one can evaluate the genetic distance between language *T* and its descendant *L* on the basis of the amount of words retained in the core lexicon of *L* from the core lexicon of *T*.

§26. This suggestion contradicts the claims made in several substantial publications of Blust. According to him, the Austronesian languages have 'evident variability in retention percent' (Blust 1993:245), and percents of words retained by sister-languages from their common ancestor vary within 20-30 percent. If correct, this completely undermines the lexicostatistical method.

Let us, however, discuss Blust's arguments.

For many years, this scholar developed a classification of the Austronesian languages. This widely accepted classification has grouped the languages in the following tree:

Austronesian Family

1. Atayalic
2. Rukai-Tsoic
3. Paiwanic
4. Malayo-Polynesian (MP):
  - i. Western Malayo-Polynesian
  - ii. Central-Eastern Malayo-Polynesian (CEMP):
    - a. Central Malayo-Polynesian (CMP)
    - b. Eastern Malayo-Polynesian (EMP):
      - South Halmahera - West New Guinea (SHWNG)
      - Oceanic (Oc) (Blust 1978; Tryon 1995).

Blust has reconstructed the  $\sigma$ -lists for main proto languages of Malayo-Polynesian: PMP, PCEMP, PEMP and POc. Comparing these reconstructed lists with the  $\sigma$ -lists of various languages, he came to the following conclusion: "Almost all CMP languages, apart from those of the Aru Islands, are lexically quite conservative, with a mean retention percent of reconstructed PMP basic vocabulary of 38.9. In other words, the typical CMP language has a high concentration of

lexical items that belong to cognate sets that are widely distributed in the Austronesian family. The SHWNG languages, on the other hand, are only moderately conservative (mean retention percent 25.6). The Oceanic languages vary widely in retention percent, from lexically rather conservative languages such as Ruga (39.5), Fijian (39.5), Trukese (37.8), Motu (36.7), Sa'a (36.2) and the Polynesian languages (ranging from about 33 to 40 percent) to lexically very innovative languages such as Jawe (19.1), Roviana (16.5), Misima (15.7), Kilivila (14.6), Teanu (10.8), Dehu (9.8), or Kaulong (5.7)" (Blust 1993,245).

Evaluating these conclusions, we need to address three main issues:

- the genetic classification of the Austronesian languages;
- the notion of 'retention rates'
- the procedure of evaluating these rates.

We cannot discuss here the whole problem of genetic classification of Austronesian (for an overview see Ross 1995), especially the question as to what extent the proposed Blust's classifications is a genetic one. To prove the genetic nature of a classification, one needs to provide conclusive evidence that all features used to justify the suggested branching are of pure genetic origin and do not represent regional convergence, borrowing or other non-genetic developments. In his classifications Blust follows the common technique of using selected innovations (phonological, grammatical, lexical) to prove the groupings. However, when dealing with an innovation, we always have a good chance that it be an externally caused and/or triggered change. Often, even for such well-known families as Slavonic, it is very difficult to prove the opposite and to demonstrate that a feature treated as an innovation is a free internally caused change. Austronesian comparative studies belong to the most developed areas of comparative linguistics, but still it is too early to believe that we can rule out the possibility of non-genetic origins of group-specific innovations, especially when such features are mainly losses or merges (see, for example Blust's list of MP innovations - Blust 1990).

The idea of retention rate, as we understand it<sup>20</sup>, can be presented as the following: one can compare  $\sigma$ -lists of languages L and its direct ancestor and count the percent of words which are genetically identical in these two lists. This percent represents the retention rate of L.

<sup>20</sup> The original unpublished work of Blust dealing with the retention rates is not available to us.

Retention rates were established for several pairs of languages (see, for example, Bergsland and Vogt 1962; Starostin 1989). All these test cases, however, are based on comparisons between recorded and well-known languages, which allows scholars to compile the  $\sigma$ -list with full confidence, choosing appropriate list-members.

It is much more difficult to compile a  $\sigma$ -list for a reconstructed language, than for a well known recorded one, as normally we do not have sufficient tools to prove that a word meets all the conditions required for a list-member: to be the main unmarked representation of an  $\alpha$ -meaning, to belong to a particular dialect and at a chronological level and so on. In such cases, a word with a wider distribution can be seen as the proper list-member, which is not always true (see, for example, discussion of 'aging' of words in Starostin 1989) and additional research is needed to support that the most widely spread modern reflexes indicate the proper list-member of the proto language's  $\sigma$ -list. With many hundreds of Austronesian languages yet to be synchronically described, it seems to be practically impossible to substantiate sufficiently such claims<sup>21</sup>.

Therefore we have solid reasons to believe that Blust's lists for various proto languages are not lexicostatistical  $\sigma$ -lists, but are lists of reconstructed forms whose reflexes in modern languages have wider distribution and are also represented in corresponding ancestor languages. This is supported by the following observation made by Blust about PMP, PCEMP, PCMP and POc: "The lexicostatistical comparison of the four protolanguages is of some limited interest. Systematic attempts to reconstruct Swadesh's 200-item test lists at various time-depths show clearly that Proto-Central Malayo-Polynesian was hardly distinct from Proto-Malayo-Polynesian (98 percent similar). A comparable relationship holds for Proto-Central Malayo-Polynesian in relation to Proto-Central-Eastern Malayo-Polynesian (96 percent). The comparison of other pairs of these protolanguages yields only moderately lower values: PCMP and PMP (94 percent), POc and PCEMP (93 percent), POc and PMP (88 percent), POc and PCMP (84 percent)". (Blust 1993, 245).

<sup>21</sup> In fact, due to the lack of developed reconstructions in Austronesian studies linguists are often dealing with 'comparisons' rather than with 'etymologies' (see Peiros, in this volume).

The percents given by Blust form the following matrix:

	PMP	PCEMP	PCMP	POc
PMP	x	98	94	88
PCEMP	96	x	96	93
PCMP	94	96	x	84
POc	88	93	84	x

Note that this matrix is based on 200-item list. For the standard list Blust's data allows us to build the following matrix:

	PMP	PCEMP	PCMP	POc
PMP	x	98	96	87
PCEMP	98	x	100	90
PCMP	96	100	x	
POc	87	90	89	x

It follows from this matrix that the reconstructed PCEMP list is identical to that of PC and nearly identical to the list for PMP. Accepting these results we have to conclude that no lexical changes have occurred over the whole period elapsed from PMP to PCEMP, which is quite improbable. It is better to assume that the matrix shows simply that forms reconstructed for PMP were preserved in Proto CEMP leaving completion of proper  $\sigma$ -lists for the time being.

If our understanding of the percents studied by Blust is correct, we can say that they, in fact, represent the extent to which languages confirm the suggested reconstructions, and they do not provide us with conclusive evidence of significant differences of retention rates among the Oceanic languages, and thus they do not contradict the theory of lexicostatistics.

§27. A lexicostatistical matrix includes sufficient data to model the evolution of a family. Percents included in it indicate directly degrees of diversity between languages, while their interpretation (not discussed here) provides us with the genetic tree of the family.

Here we cannot discuss the formal procedure of matrix interpretation (see Peiros and Starostin, in progress) which is based on the assumption that an

evolutional tree can be obtained only from the whole matrix, rather than through analyses of percents revealed by individual pairs of languages. This procedure is used in STARLING, a software package designed by Starostin.

## LITERATURE

- Alpher, B. & Nash, D. (forthcoming) *Lexical replacement and cognate equilibrium in Australia*.  
45 pp.
- Bergsland and Vogt, 1962, *On the validity of glottochronology*.  
In Current Anthropology 3: 111-153.
- Biggs, B., 1991, *A Linguist revisits the New Zealand Bush*.  
In Pawley, A., ed., Man And a half, Auckland: The Polynesian Society, pp. 67-72.
- Biggs, B., MS POLLEX (computer print-out, University of Auckland).
- Blust, R., 1978, *Eastern Malayo-Polynesian: a subgrouping argument*.  
In S. Wurm and L. Carrington, eds., Second International Conference on Australian Linguistics: proceedings. PL, C-39: 181-234.
- Blust, R., 1981, *Variation in retention rate among Austronesian languages*.  
Paper presented to the Third International Conference on Austronesian Linguistics, Bali, Indonesia (not seen).
- Blust, R., 1990, *Patterns of sound change in the Austronesian languages*.  
In Baldi, P. ed. Linguistic change and reconstruction methodology.  
Berlin, New York: Mouton de Gruyter, 231-263.
- Blust, R., 1993, *Central and Central-Eastern Malayo-Polynesian*.  
In Oceanic Linguistics, 32/2: 241-293.
- Dixon, R., 1980, *The languages of Australia*.  
Cambridge: Cambridge University Press.
- Dixon, R., 1983, *Nyawaygi*.  
In Dixon R. and B. Black, eds., The Handbook of Australian Languages. Vol. 3.  
Canberra: Australian National University Press, 430 -531.
- Guy, J.B.M., 1981, *Glottochronology without cognate recognition*.  
In Pacific Linguistics, B-79.
- Mel'chuk I., 1988, *Dependency Syntax: theory and practice*.  
N.Y.: State Univ. of New York Press.
- Nichols, Johanna, 1996, *The comparative method as heuristic*.  
In M. Durie and M. Ross. The comparative method reviewed.  
Oxford: Oxford University Press, pp. 39 - 71.



Peiros, I., 1989, *Languages and Sociolanguages*.

Paper read in the Institute of Ethnography, Soviet Academy of Sciences, Moscow.

Peiros, I., 1995, *Proto Katuic Comparative Dictionary*.

In Pacific Linguistics C-132. vi+192 pp.

Peiros, I., 1996, *The Vietnamese etymological dictionary and 'new' language families*.

In Pan-Asiatic linguistics. Mahidol University: Salaya, vol 3 pp. 883-895.

Peiros, I., 1998, *Linguistic Prehistory of Southeast Asia* (396 pp.).

In Pacific Linguistics, scheduled for publication in 1998.

Peiros, I., MS *Proto Palaung-Wa reconstruction and Comparative dictionary*.

Peiros, I., and S.Starostin, 1996, *A Comparative Vocabulary of Five Sino-Tibetan languages*. Fasc.1-6.

Melbourne: Department of Linguistics and Applied Linguistics.

Peiros, I., and S. Starostin, in progress, *Lexicostatistics revisited*.

Ross, M., 1995, *Some current issues in Austronesian linguistics*.

In Tryon, D., ed., Comparative Austronesian Dictionary, Part 1: 45-120.

Berlin, New York: Mouton de Gruyter.

Simpson, J. *How Warumungu people express new concepts*.

In Language in Central Australia, 4: 12-25.

Starostin, S. 1989, *Comparative Linguistics and Lexicostatistics*.

In Lingvističeskaja rekonstruktsija i drevnejšaja istorija Vostoka (Linguistic reconstruction and ancient history of the Orient, in Russian).

Moscow, Nauka, part 1: 3 - 39.

Tryon, D.T., ed., 1995, *Comparative Austronesian Dictionary*. Parts 1-4.

Berlin, New York: Mouton de Gruyter.



# INDICES

## INDEX OF AUTHORS

Birnbaum, Henrik, 111  
 Guardans, Ramon, 201  
 Gurevich, Naomi, 119  
 Jaxontov, Sergei E., 51  
 Manaster-Ramer, Alexis, 95  
 Parkinson, James B., 105

Pejros, Ilia I., 259  
 Smiljanić, Rajka, 145  
 Starostin, Sergei A., 3, 61  
 Sverdrup, Harald, 169, 201  
 Vovin, Alexander, 67

## CITATION INDEX

Albert, Roy, 67<sup>f</sup>  
 Alpher, Barry, 300  
 Ambrosini, R., 202  
 Ammerman, Albert J., 171  
 Anderson, J.M., 202  
 Anttila, Raimo, 111  
 Arapov, M., 9, 27  
 Bergsland, Knut, 11, 13, 58, 112, 310  
 Biggs, B., 265, 313<sup>f</sup>  
 Bird, R. Byron, 171  
 Blust, Robert A., 308, 309<sup>ff</sup>  
 Bonfante, G., 202  
 Bonfante, Larissa, 202  
 Cavalli-Sforza, Luigi Luca, 171, 176, 187, 189  
 Chang, T., 176, 185  
 Chrétien, C. Douglas, 23<sup>f</sup>, 56-57, 59  
 Comrie, Bernard, 77  
 Cristofani, M., 202  
 Dixon, R.M.W., 299, 300  
 Doerfer, Gerhard, 67, 77  
 Dolgopolsky, Aharon B., 95-101, 120  
 Durie, Marc, 261<sup>f</sup>  
 Dybo, Vladimir Antonovič, 105, 108  
 Dyen, Isidore, 21, 23<sup>f</sup>  
 Daukjan, Gevork Beglarovič, 56  
 Embleton, Sheila, 13<sup>f</sup>  
 Fagan, Brian, 176  
 Fodor, István, 14, 111  
 Forsyth, K., 205  
 Gleason, G., 57  
 Greenberg, Joseph H., 95, 96, 100, 119<sup>ff</sup>, 135, 145<sup>ff</sup>  
 Guardans, Ramon, 201<sup>ff</sup>  
 Guryčeva, M., 16<sup>f</sup>  
 Guy, J.B.M., 281<sup>f</sup>  
 Hattori, Shirō, 13<sup>f</sup>, 24<sup>f</sup>  
 Herz, M., 9, 27

Hock, Hans H., 120, 150<sup>f</sup>  
 Howells, W.W., 176  
 Hubschmied, Johannes, 202  
 Illič-Svityč, Vladislav Markovič, 3, 4, 96, 105  
 Jackson, K.H., 202  
 James, A., 21  
 Janhunen, Juha, 77  
 Jaxontov, Sergei E. 11, 25<sup>f</sup>  
 Kaiser, Mark, 109  
 Kalectaca, Milo, 67<sup>f</sup>  
 Kaufman, Terrence, 98  
 Kretschmer, P., 202  
 Kuper, R., 183  
 Lamb, Sydney M., 114  
 Lamberg-Karlovsky, C.C., 181  
 Langacker, Ronald W., 98  
 Larichev, V. 176  
 Locker, E., 205  
 MacAlister, R.A.S., 202, 205  
 Manaster-Ramer, Alexis, 67<sup>f</sup>, 95, 100  
 Manczak, W., 111  
 Mel'chuk, Igor A., 279, 293  
 Menges, Karl H., 77  
 Menozzi, P., 184  
 Militarev, Aleksander Yuri, 47  
 Miller, Roy Andrew, 77  
 Miller, Wick R., 67<sup>f</sup>  
 Mountain, J., 184  
 Müller, H., 16<sup>f</sup>  
 Murayama, Shichirō, 77  
 Nash, D., 300  
 Nichols, Joanna, 281<sup>f</sup>  
 O'Neil, W., 12  
 Pallotino, M., 202  
 Pattison, W., 202  
 Pedersen, Holger, 96  
 Pejros, Ilia I., 281, 304

- Peter, Steven J., 120, 135, 147  
 Piazza, A., 184  
 Pulgram, Ernst, 114  
 Renfrew, Colin, 170, 176  
 Rhys, John, 202, 205  
 Ringe, Donald R., 121<sup>ff.</sup>, 147  
 Rix, H., 202, 204, 206  
 Roman del Cerro, J.L., 204, 212  
 Ross, M., 309  
 Ruhlen, Merritt, 185  
 Scharf, J., 212  
 Schuhmacher, S., 202, 205  
 Shevoroshkin, Vitaly V., 61<sup>f</sup>  
 Shaul, David L., 67<sup>f</sup>  
 Sidwell, Paul, 95  
 Simpson, J., 265  
 Sirk, Y., 47  
 Starostin, Sergei A., 77, 111, 120, 261, 304, 310, 312  
 Strahlenberg, Philipp Johann von, 95  
 Street, John, 77  
 Sverdrup, Harald, 201<sup>ff.</sup>  
 Swadesh, Morris, 8, 11, 13<sup>f</sup>, 25<sup>f</sup>, 51, 59, 120, 296  
 Tovar, Antonio, 212, 216  
 Tryon, Darrell T., 301, 309  
 Untermann, Jürgen, 202  
 van der Merwe, N., 21  
 Vennemann, Theo., 202, 205, 212, 216<sup>f</sup>  
 Villar, F., 202  
 Vogt, Hans, 11, 13, 58, 112, 310  
 Vovin, Alexander, 77  
 Wessén, E., 112  
 Zuravlev, A.F., 114

Note: Footnotes are indicated by superscripts. E.g., 'Dyen, Isadore, 21, 23<sup>f</sup>' means Isadore Dyen is referenced in the main text of page 21, and in a footnote of page 23.

# INDEX OF SUBJECTS

- accidental resemblances, 5, 120, 122, 133<sup>ff.</sup>, 277  
 agriculturalists (neolithic), 171  
 Ainu, 4  
 Altaic (*family*), 77, 82, 106  
 Amerind (*family*), 96, 100, 189  
 Austric, 189  
 basic lists (BL) of words, 9-25, 28-35, 38-44, 54, 67, 68-71, 73-76, 78-81, 82-86, 87-90, 115, 120, 137-139, 139-141, 155-159, 159-163, 212-215, 291  
 Basque, 4, 181, 186, 208, 209  
 Bokmal (Riksmal), 11, 112  
 borrowings, 124, 134, 145, 150, 260, 290  
 Capsian culture, 183  
 Catal Hüyük, 175  
 census, 181<sup>f</sup>  
 chance resemblance, 122, 277. *see* accidental resemblances  
 change, 265<sup>ff.</sup>  
 causes of, 264  
 free, 269<sup>ff.</sup>  
 in meanings, 105<sup>ff.</sup>, 265  
 in phonetics, 105<sup>ff.</sup>  
 triggered, 269<sup>ff.</sup>  
 China, 176  
 Han (dynasty) census, 181<sup>f</sup>  
 Chinese, 13<sup>f</sup>  
 chronology, linguistic, 4, 7, 51<sup>ff.</sup>  
 mathematics of, 9, 19, 22-23, 45  
 classification (language), 95, 119, 217  
 cognate divergence rate, 8, 11-12  
 community, 263, 265<sup>ff.</sup>  
 crystallization, 264  
 disintegration, 264  
 information network, 261, 264  
 comparative historical analysis, 47, 105  
 comparative linguistics, 3, 4, 53, 61<sup>f</sup>, 67, 263  
 comparative method, 91, 145  
 reconstruction, 283  
 comparison, long-range, 61, 67  
 computer technology, 64

- consonants
  - unused, 215-216
- Coptic-Berber, 183
- core lexicon, 282, 294*f.*
- core meanings (s), 265, 294
- Cro-Magnon, 183, 187
- Cupeño (*of* Utu-Aztec), 97*ff.*
- Dano-Norwegian (*see* Bokmal or Riksmål)
- diachronic sound changes, 145
- Dravidian (*family*), 106, 181
- driving force (for change), 170
- Dryas cold periods, 180
- ecological systems, 169
- Egyptian (ancient), 105
- English, 129*ff.*, 137*ff.*, 150*ff.*
- Etruscan, 181, 186*f.*, 201*ff.*, 210, 217
- etymological dictionaries, 164*f.*
- etymological statistics, 26, 47, 105, 112, 114
- etymology, 124, 133, 145
- evolutional tree (of languages), 270, 278
- false friends, *see* accidental resemblances
- Faroese, 12
- Finnish, 129*ff.*, 139*ff.*, 150*ff.*
- French, 14<sup>f</sup>, 16, 28*ff.*, 38*ff.*
- genetic links (languages), 61, 154
- genetic tree (of languages), 6, 65-66, 276*ff.*
- genetics (population), 176
- German, 14<sup>f</sup>, 28*ff.*, 38*ff.*, 113, 129*ff.*, 139*ff.*, 150*ff.*
- glaciation, 216
- gloss, 68
- glottochronology, 4, 8, 51, 59, 111, 261<sup>f</sup>, 285
  - divergence rate constant, 11, 14
  - mathematics of, 9, 19, 22-23, 45, 52
  - probabilistic nature of, 10
  - root, 26, 45, 115, 120
- glottogenesis, 266
- grammar, 56
- Grimm's law, 124, 150
- Hebrew, 105
- Hindhi, 129*ff.*, 139*ff.*, 150*ff.*
- historical linguistics, 3
  - mass comparison, 119*ff.*, 135*f.*, 145*ff.*
  - methodology, 7, 119-135
  - homelands (linguistic), 111, 176*ff.*
  - horses (domestication), 185
  - hunter-gatherers (mesolithic), 171
  - hydronymy, 216
- Iberian, 181, 186*f.*, 201*ff.*, 209, 232-238
- Icelandic, 11, 12, 112
- ideophones, 5
- index of retention, 52
- Indo-European (IE), 38*ff.*, 101, 106, 123
- Japanese, 13<sup>f</sup>, 67*ff.*, 77*ff.*
- Jericho, 175, 176, 183
- Kartvelian (*family*), 106
- Koptic, *see* Coptic
- Korean, 77
- language (j, iazyk), 262*f.*, 265*ff.*
  - classification, 63, 95, 276
  - modelling, 287
  - continuity, 270
  - daughter, 270
  - descendant, 270
  - divergence, 3
  - ergativity, 215
  - expressive means, 287
  - family, 270, 282, 284
  - non-specific relationship, 298
  - reconstruction, 4, 279*f.*
  - sister, 274
  - sociolanguage, 260*f.*
  - specific relationship, 272, 298
  - splits, 114
- LANGUAGE model (math.), 172*ff.*, 189
- languages
  - Afro-Asiatic (AA, *or* S-H), 105, 183
  - Ainu, 4
  - Altaic, 77, 106
  - Amerind, 96, 100, 189
  - Austic (*group*), 185, 189
  - Australian Aboriginal (*group*), 291
  - Austronesian (*group*), 300
  - Austro-Asiatic, 185
  - Basque, 4, 181, 186*f.*, 208, 209
  - Burushaski, 185
  - Camunian, 203, 208, 210

- Caucasian (*group*), 181  
 Chinese, 13<sup>1</sup>  
 Cupeño (*of* Utu-Aztec), 97ff.  
 Cushitic, 183  
 Danish, 112  
 Dene-Sino-Caucasian, 177, 187, 189  
 Dravidian, 106, 181  
 Egyptian (ancient), 105  
 Elamitic (*group*), 181  
 Elymian, 204, 208, 210  
 English, 14<sup>1</sup>, 28ff., 38ff., 112,  
 129ff., 137ff., 150ff.  
 Etruscan, 181, 186f., 201ff., 210,  
 217, 220-232  
 Faroese, 12  
 Fijian, 309  
 Finnish, 129ff., 139ff., 150ff.  
 French, 14<sup>1</sup>, 16, 28ff., 38ff.  
 German, 14<sup>1</sup>, 28ff., 38ff., 113,  
 129ff., 139ff., 150ff.  
 Greek (ancient), 36f.  
 Guarjio (*of* Utu-Aztec), 97  
 Hebrew, 105  
 Hindi, 129ff., 139ff., 150ff.  
 Hopi (*of* Utu-Aztec), 67  
 Hungarian, 112  
 Iberian, 181, 186f., 201ff., 209, 232-  
 238  
 Icelandic, 11, 12, 112  
 Indo-European (IE), 38ff., 101, 106  
 Itturan, 209, 241  
 Japanese, 13<sup>1</sup>, 67ff., 77ff.  
 Japanese (Old), 71f., 82  
 Kartvelian, 106  
 Landsmal, *see* Nynorsk  
 Latin, 36, 114  
 Lemnian, *see* Lemno-Pelasgian  
 Lemno-Pelasgian, 204, 208, 210, 238  
 Lepontic, 204, 208, 211, 238f.  
 Ligurian, 204, 240  
 Lithuanian, 28ff., 38ff.  
 Malayo-Polynesian, 308f.  
 Marathi, 129ff., 139ff., 150ff.  
 Miao-Yiao, 185  
 Nahuatl (*of* Utu-Aztec), 97ff.  
 Na Dene, 180  
 North Pikene, 205, 210, 239f.  
 Norwegian, 11, 112  
 Nostratic, 3, 4, 6, 95, 96, 105-109,  
 181, 189  
 Nuragic, 205, 240  
 Nyawaygi, 300  
 Nynorsk (New Norwegian), 112  
 Paleoeuropean, 205, 217, 241  
 Pictish, 181, 186f., 201ff., 209, 242  
 Pikene (North), 205, 210  
 Polish, 36  
 Proto-Afro-Asiatic, 176  
 Proto-Austro-Asiatic, 181  
 Proto-Caucasian, 176  
 Proto-Dene-Sino-Caucasian, 179  
 Proto-Elamo-Dravidian, 176, 181  
 Proto-Indoeuropean, 176  
 Proto-Japanese (PJ), 72ff.  
 Proto-Ugric-Altaic, 176  
 Proto-Uto-Aztec (PUA), 72ff., 97,  
 98  
 Rätian, 201ff., 210, 217, 244  
 Riksmal, 11, 112  
 Rumanian, 14<sup>1</sup>, 18  
 Russian, 28ff., 38ff., 116, 129ff.,  
 137ff.  
 Scandinavian, 12  
 Semitic, 101, 183  
 Semito-Hamitic (S-H, *or* AA), 105  
 Serbo-Croatian, 150ff.  
 Sino-Caucasian, 189  
 Slavonic (Old Church), 116  
 Spanish, 14<sup>1</sup>, 17  
 Sumerian, 4  
 Swedish, 112  
 Tarahumara (*of* Utu-Aztec), 99  
 Tartessian, 201ff., 212, 248  
 Tübatulabal (*of* Uto-Aztec), 96f.,  
 99  
 Turkic (Old), 82  
 Turkish, 77ff.  
 Uralic, 105  
 Uto-Aztec (UA), 96  
 Vedic, 36f.  
 Yenesseian, 180
- lexemes  
   etymological analysis, 145  
   stability of, 95  
 lexical shifting, 105-109, 265  
 lexicon, 275  
 lexicostatistics, 13, 47, 63, 111, 120,  
 259ff.  
   data base, 297  
   basic postulates, 287  
   diagnostic parameters, 286, 290  
   lexicostatistical matrix, 298f.  
   lexicostatistical tables, 294f.  
   shortcomings, 25  
   σ-list, 290, 291f.

- list members, 290
- linguistics
  - comparative, *see* comparative linguistics
  - comparative historical, 4
  - dating, *see* glottochronology
  - historical, *see* historical linguistics
  - LANGUAGE model, 172ff.
    - archaeological postulates, 175
    - historical checking, 175
    - myth about lost similarity, 288
  - literary norms, 12
  - loan words, 5, 111
  - long-range comparison, 61, 67
    - statistical methods, 63
- Manchu-Tungus, 77
- Marathi, 129ff., 139ff., 150ff.
- mass comparison, 119ff., 135f., 145ff.
- mesolithic population, 171
- migration, 265
- mock data, 121, 124
- Mongolian, 77
- morphemes, 281
  - etymological analysis, 145
  - stability of, 95
- morphology, 275
- Nahuatl (*of* Uto-Aztecan), 97ff.
- Natufian culture, 175
- neolithic population, 171
- Norwegian, 11, 112
- Nostratic, 3, 4, 6, 95, 105-109, 181, 189
- nursery language, 124, 134, 145, 150
- Ogham script, 205
- Old Japanese, 71f., 82
- Old Turkic, 82
- onomatopoeia, 124, 134, 150
- Palestine, 175
- phonetic changes, 105-109, 113
  - a method for evaluating, 108f.
- phonetic correspondences, 71f., 76f., 81f., 86, 91, 145, 154, 278
- phonology, 56
- Pictish, 181, 186f., 201ff., 209
- population dynamics, 169
- population genetics, 176
- preprotolanguages, 111
- pronoun
  - animate, 231
  - class e-, 231
  - class i-, 231
  - inanimate, 231
  - neutral, 231
- proto-languages, 4, 111
  - ancestral, 4
  - Proto-Altaic, 90f.
  - Proto-Japanese, 72ff., 87ff.
  - Proto-Turkic, 87ff.
  - Proto-Uto-Aztecan, 72ff., 97, 98
- reconstruction (proto-language), 4, 62, 119, 136, 145, 283
- remote comparison, 4
- resemblance
  - accidental, 5, 120, 122, 133ff., 282
  - chain reaction, 126
  - suspected, 125f., 128ff.
- rigorous methodology, 3, 145f., 154
- Riksmal (Bokmal), 11, 112
- root glottochronology, 26, 45, 115, 120
- Russian, 28ff., 38ff., 116, 129ff., 137ff.
- Scandinavian, 12
- Semito-Hamitic (S-H, *or* AA), 105
- Serbo-Croatian, 150ff.
- Slavic, 113
- Slavonic (Old Church), 116
- sociolanguage, 262f.
- sound correspondences, *see* phonetic correspondences
- STARLING (software), 304
- statistics, 63
- Sumerian, 4
- Swadesh constant, 14
- Swadesh method, 55, 112
- synonyms, 123<sup>f</sup>, 127
- toponyms, 212, 216
- Turkic, 77
- Turkish, 77ff.
- Uralic (*family*), 106
- Uto-Aztecan (UA), 96
- vocabulary, 56
- words (selected)
  - aging, 302
  - most stable, 25<sup>f</sup>, 120

- \*aiw-, 32
- \*Haster-, 43
- \*awes-, 34
- \*bheu-, 31
- \*bhreu-, 38
- \*dekm-, 32
- \*derw-, 33
- \*dheu-, 39
- \*dhreugh-, 39
- \*duM-, \*dwo(u)-, 28, 44
- \*ed-, 39
- \*eg(h)om-, 41
- \*er(t)-, 39
- \*es-, 30
- \*gel-, 38
- \*g<sup>w</sup>el-, 41
- \*gnM-, 41, 151
- \*kap-, 40
- \*kerd-, 40
- \*kern-, 40
- \*keuk-, 34
- \*k<sup>w</sup>o-, 29
- \*leg<sup>h</sup>-, 41
- \*leubh-, 41
- \*men-, 42
- \*m nes-, 41
- \*mon-, 41
- \*nas-, 42
- \*ne-, 31, 42
- \*new-, 42
- \*newo-, 30
- \*no-(s), 31
- \*<sup>(o)</sup>nogh<sup>w</sup>-, 42
- \*nok<sup>w</sup>t-, 42
- \*<sup>(e)</sup>nomn-, 42
- \*oi-n-, 42
- \*ok<sup>w</sup>-, 39
- \*ous-, 39
- \*peHwMr-, 39
- \*peisk-, 39
- \*penk<sup>w</sup>e-, 35
- \*per- (front side), 5, 29
- \*rek-, 42
- \*reudh-, 42
- \*s-(men)-, 43
- \*sek<sup>w</sup>-, 43
- \*smeug-, 43
- \*st-, 43
- \*stei-, 43
- \*swel-, 43
- \*to-, 28, 43
- \*tong-, 35
- \*we-, 30, 44
- \*wed-, 44
- \*weid-, 28
- \*wes-, 29, 34, 41
- \*wer-, 35
- \*werg-, 28
- \*wiro-, 30
- a (once, any), 28
- altar, 246
- are, 30
- be, 31, 220, 232, 242
- belly, 21
- blood, 38
- bone, 38
- bull, 228
- burn, 38
- cereal, 234
- children, 242
- cloud, 21
- cold, 38
- consecrate, 226
- darkness, 235
- dead, die, 39, 68, 73, 78, 83, 87, 224, 225, 238, 242, (243), 246(2)
- dog, 240
- dry, 39
- ear, 21, 39, (62), 137, 140, 141, 156, 161
- earth, 39
- East, 34
- eat, 39
- (for)ever, 31
- eye, 21, 39, 68, 73, 78, 83, 87, 137, 156, 161
- father, 221, 232, 242, 244
- fingerail, *see* nail
- fire, 12, 39, 137, 139, 156, (161), 223, 229
- fish, 39
- five, 35, 62
- full, 40
- God, 213, (218), 220, 233, (238), 244
- govern, 228
- grain, 234
- grandfather, 226
- grave, 213, 251
- hand, 40, 101, 215, 232
- head, 40, 213
- hear, 62
- heart, 40, 63, 69, 74, 79, 84, 88
- here, 237



- high(er), 34, 213, 220, 229, 237, 238, 241  
 hill, 233, 242  
 horn, 40, 138, 140, 157, 296  
 house, 213, 233, 240, 244  
 I, 21, 35, 41, 69, 74, 79, 84, 88, 100, 214, 218, 219(2), 225, 231, 236, 243, 246  
 in, 28  
 is, 30  
 kill, 41, 138, (140, 141), 157, (162), 232  
 know, 31, 41, 138, 140, 141, 151, 157, (162)  
 land, 213, 241  
 law, 226  
 leaf, 12, 41  
 lie, 41  
 lion, 224  
 louse, 41, 69, 74, 79, 84, 88  
 man, 12, 41, 138, 214, 219, 222, 226, 232, 234, 243, 244, 247  
 moon, 12, 41, 228, 242, 300  
 mother, 221, 244  
 mountain, 235, 240, (247), 249  
 mouth, 42, 213  
 nail, 42, 70, 75, 79, 84, 89  
 name, 42, 70, 75, 80, 84, 89  
 negative, *see* no(t)  
 new, 30, 42, 138, 140, 141, 158, 163  
 night, 42  
 no(t), 31, 42, 70, 75, 80, 85, 222  
 nose, 42, 125, 138, 140, 141, 158, 163  
 offer(ing), 228, 232  
 one, 42, 228, 230, 233, 237  
 our, us, 31  
 people, 233  
 person, 223, 293  
 rain, 42, 138  
 red, 42, 138  
 river, 224, 235, 236, 241, 245  
 rock, 235  
 root, 33, 42, 138  
 sacred, (220)  
 sacrifice, 223, 233  
 sea, 220  
 see, 43, 137, 157, 223  
 seed, 81  
 sheep, 226  
 sleep, 43  
 smoke, 43  
 son, 222, 237  
 star, 43  
 stem, 33  
 stone, 43, 81, 213, 221, 222, 235, 235, 240(2), 241(2), 242(2), 245, 301, 303<sup>f</sup>  
 sun, 21, 43, 228, 232  
 tail, 21  
 tear (noun), (235)  
 temple, 227  
 ten, 32, 226, 230, 236, 237  
 the (this, that), 28, 43, (222, 224, 227, 231)  
 tongue, 44, 71, 76, 81, 85, 90  
 tooth, 44, 71, 76, 81, 85, 90  
 tree, 33, 44  
 two, 28, 44, 62, 71, 76, 81, 85, 90, 138, 140, 141, 159, 163, 230, 232, 233, 237  
 valley, 233, 241  
 was, 29  
 water, 44, 71, 76, 81, 86, 90, 213, 215, 225, 235, 237, 240, 241(4)  
 we, 30, 44  
 well (water), 226, 235, 241  
 West, 34  
 who, what, 29, 44, 71, 76, 81, 86, 90  
 wind, 139, 155, 160  
 wine, 229, (248)  
 woman, 44  
 wood, 33  
 word, 35  
 work, 28  
 world, 30  
 year, 32  
 you, 21, 44, 71, 76, 81, 86, 90, 237  
 you (2nd *p. sing. only*), 96-101, 235, 245  
 wheat harvesting, 175  
 Wurm-Weichsel glacial maximum, 179<sup>f</sup>  
 Zagros Mountains, 176



# Association for the History of Language Monographs

OUT NOW!

## *Professing Koernerian Linguistics*

A Selection of Papers and Reviews presented in honour of  
Professor E. F. K. Koerner

Published May 1998  
ISBN 0 7340 1356 6  
Paperback xiv+148 pages  
USD \$25.00 (postage paid)

## *From Neanderthal to Easter Island*

A tribute to, and a celebration of, the work of  
W. Wilfried Schuhmacher. Presented on the occasion of his 60th Birthday

Published June 1999  
ISBN 0 9577251 0 8  
Paperback x+165 pages  
USD \$25.00 (postage paid)

*Contact AHL by mail at:*

LPO Box A22, Australian National University, ACT 0200 Australia

*or visit our web page at:*

<http://www.lexicon.net./opoudjis/Work/ahl.html>.

*Не можем молчать!*